From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

FLEXIBLE PARAMETRIC MODELS FOR CANCER PATIENT SURVIVAL: LOSS IN EXPECTATION OF LIFE AND FURTHER DEVELOPMENTS

Hannah Bower

Karolinska Institutet

Stockholm 2018

Published by Karolinska Institutet.
Printed by E-Print AB.

ISBN 978-91-7676-925-6

Karolinska Institutet

Department of Medical Epidemiology and Biostatistics

Flexible Parametric Models for Cancer Patient Survival: Loss in Expectation of Life and Further Developments

AKADEMISK AVHANDLING

som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i hörsal Hillarp, Retzius väg 8, Karolinska Institutet, Solna

Fredagen den 23 mars 2018, kl 09.00

av

Hannah Bower

Huvudhandledare:
Paul C Lambert, Professor
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics
and
University of Leicester, UK
Department of Health Sciences
Biostatistics Research Group

Bihandledare:
Paul W Dickman, Professor
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Therese M-L Andersson, PhD
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Mats Lambe, Professor
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics
and
Regional Cancer Center Uppsala Örebro

Fakultetsopponent:
Angela Mariotto, PhD
The National Cancer Institute, US
Division of Cancer Control and Population Sciences

Betygsnämnd:
Nicola Orsini, Associate Professor
Karolinska Institutet
Department of Public Health Sciences

Max Petzold, Professor
University of Gothenburg
Department of Public Health and Community Medicine

Sara Hägg, Associate Professor
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Stockholm 2018

Abstract

Population-based cancer studies can contribute to a better understanding of cancer patient survival. The general aim of this thesis was to develop and apply statistical methods for population-based cancer studies to ensure understanding in this area.

In any analysis setting, it is important that the statistical methods are appropriate. Since non-proportional hazards are common in population-based data where follow-up is long, **Study I** assessed the ability of flexible parametric models (FPMs) in capturing time-dependent effects in a simulation setting. This study also attempted to determine whether the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) were more appropriate in selecting a good-fitting model. Results indicated that bias was small for estimated survival proportions, hazard rates and log hazard ratios when degrees of freedom were selected using guidance from the AIC or BIC. Neither the AIC nor the BIC constantly outperformed the other. We concluded that FPMs accurately capture time-dependent effects but that users should perform appropriate sensitivity analyses.

There have been large changes in treatments for chronic myeloid leukaemia (CML) patients over time; most notably is the introduction of imatinib mesylate, a targeted therapy, in 2001. Studies have shown that relative survival of CML patients greatly improved after the introduction of this treatment. Since CML is a chronic disease, quantifying survival in terms of life expectancy is highly relevant. In **Study II** we aimed to quantify life expectancy improvements using the loss in expectation of life measure. Results indicated that patients diagnosed in 2013 would be expected to lose less than three of their remaining life years due to their CML diagnosis. We concluded that the life expectancy among CML patients is approaching that seen in the general population and improvements over time are largely due to imatinib mesylate and allogeneic stem cell transplantation.

Population-based mortality rates are often used as a proxy for the mortality rate a diseased population would have experienced had they been disease-free; these are used in relative survival analysis and when calculating standardised mortality ratios. Population-based mortality rates are commonly available by age, calendar year and sex which might limit analyses to these factors. In **Study III** we described methodology to adjust population mortality rates by additional variables, such as socioeconomic status, that are available in a control population. We presented both Poisson and flexible parametric methods and found that both methods estimate similar mortality rates by socioeconomic status. Adjusting for the additional uncertainty associated with the methodology presented made little difference to five-year relative survival of breast cancer patients.

Socioeconomic status is known to affect the survival of breast cancer patients. Stage at diagnosis of breast cancer is also known to have a strong effect on survival and the distribution of stage at diagnosis often differs by socioeconomic status. In **Study IV** we aimed to quantify survival differences between education groups as a proxy for socioeconomic status. We estimated that 572 life years could be saved, and 25 deaths could be postponed five years beyond diagnosis, if differences between education group in the breast cancer stage-distribution could be removed in three regions of Sweden. If differences between education groups in stage-specific breast cancer survival could be removed, we estimated that 692 life years could be saved, and 27 deaths postponed five years beyond diagnosis.

In conclusion, this thesis reassures users of FPMs of their performance under non-proportional hazards. Secondly, methodology is described to further adjust expected mortality rates which could aid researchers in estimating survival measures by additional, non-standard variables. Finally, this thesis provides a greater understanding of the probable prognosis and the burden of CML and breast cancer diagnoses via the loss in expectation of life and other similar measures.

List of publications

I. Bower H, Crowther MJ, Rutherford MJ, Andersson TM-L, Clements M, Liu X-R, Dickman PW, and Lambert PC. **Capturing simple and complex time-dependent effects using flexible parametric survival models: a simulation study.** (*Submitted*)

II. Bower H, Björkholm M, Dickman PW, Höglund M, Lambert PC, and Andersson TM-L. **Life expectancy of patients with chronic myeloid leukemia approaches the life expectancy of the general population.** *Journal of Clinical Oncology* 2016;34(24):2851-7 https://doi.org/10.1200/JCO.2015.66.2866

III. Bower H, Andersson TM-L, Crowther MJ, Dickman PW, Lambe M, and Lambert PC. **Adjusting expected mortality rates using information from a control population: an example using socioeconomic status.** *American Journal of Epidemiology* 2017, https://doi.org/10.1093/aje/kwx303

IV. Bower H, Andersson TM-L, Syriopoulou E, Rutherford MJ, Lambe M, Dickman PW, and Lambert PC. **Quantifying the potential life years saved for breast cancer patients in Sweden if differences between education groups could be eliminated.** (*Manuscript*)

These articles are referred to by their roman numerals throughout, and are presented in full at the end of this thesis.

Related publication

- Bower H, Crowther MJ, Lambert PC. **strcs: A command for fitting flexible parametric survival models on the log-hazard scale.** *The Stata Journal* 2016;16(4):989-1012

Contents

List of abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BCBaSe	Breast Cancer Database Sweden
CC	Complete Case
CI	Confidence Interval
CML	Chronic Myeloid Leukaemia
df	Degrees of freedom
FPM	Flexible Parametric Model
GLY	Gain in Life Years
HR	Hazard Ratio
IM	Imatinib Mesylate
LEL	Loss in Expectation of Life
MI	Multiple Imputation
NBHW	National Board of Health and Welfare (Socialstyrelsen)
PELL	Proportion of Expected Life Years Lost
PD	Postponable Deaths
RCS	Restricted Cubic Spline
RQRBC	Regional Breast Cancer Clinical Quality Registers
SES	Socioeconomic Status
SCR	Swedish Cancer Registry
SKL	Symmetrised Kullback-Leibler distance
TLYL	Total Life Years Lost
YLL	Years of Life Lost

1. Introduction

It is well known that most of us will be affected, either directly or indirectly, by some form(s) of cancer in our lifetime. In Sweden, and globally, the incidence of cancer is (generally) increasing [1]. Whilst screening and other improvements in the detection of cancer can contribute to this, the main reason for the increase in incidence is population aging. Malignant tumours remain the largest cause of death in Sweden, second only to cardiovascular diseases [2].

For these reasons, it is essential that research continues to focus on understanding the causes, treatments and potential cures for cancer. At the same time, it is important to determine the effectiveness of the healthcare system in terms of cancer care, and this is what population-based cancer studies aim to do. Such studies can, for example, evaluate temporal trends in cancer patient survival and evaluate the burden of cancer on a population level to provide guidance for healthcare policies and plans for cancer control whilst providing insight into the probable prognosis of patients [3, 4].

Population-based cancer studies use population-based data to quantify survival and other measures in cancer patients from a certain geographical area. Many population-based cancer studies are based on data from cancer registries that collect information on diagnoses of cancer within whole nations. Three important measures are frequently reported from such studies: incidence rates, survival proportions and mortality rates [5, 6]. Incidence rates measure the number of new cases of a disease within a population during a specific time period. To understand the probability of survival at a given time point within a group of patients, survival proportions are estimated, whilst mortality rates are death rates calculated for the whole population. Each of these measures describe different elements of the effect of cancer in a population and together they provide an evaluation of the impact and control of cancer in a population. In this thesis, focus is on the cancer patients and quantifying their survival.

Using data from Swedish registries, we quantify the impact cancer diagnoses have on the population and estimate the likely outlook for patients who are diagnosed with cancer using sophisticated statistical methods. Statistical methods are a vital element in all research areas including population-based cancer studies; without correct statistical methods and inference, results will likely be false and misleading. The work in this thesis focuses on assessing, extending and applying sophisticated statistical models that quantify the impact cancer has on a population.

2. Aims of this thesis

This doctoral project is centered around statistical methods involved in the analysis of population-based cancer studies using flexible parametric survival models. Particular focus is on applications of the loss in expectation of life measure to encourage understanding of cancer-patient survival.

More specifically, the aims were:

- To assess the ability of restricted cubic splines in flexible parametric models in capturing time-dependent effects and determine whether one of the standard information criterion used for the selection of degrees of freedom should be preferred (Study I).
- To estimate the loss in expectation of life in Swedish chronic myeloid leukaemia patients (Study II).
- To investigate and compare methods which extend population-based expected mortality rates to adjust for variables other than age, sex and year by using information available from a control population (Study III).
- To apply methods developed in Study III to be able to estimate the loss in expectation of life by highest educational achievement in Swedish breast cancer patients (Study IV).
- To quantify the total life years lost and postponable deaths of Swedish breast cancer patients if differences between education groups could be removed in terms of differences in stage-distribution and stage-specific relative survival (Study IV).

3. Background

3.1 Cancer

3.1.1 What is cancer?

Statistical methods are the focus of this thesis, but some basic understanding of cancer is necessary. Simply stated, cancer is a disease which is defined by abnormal cell growth. Most cells in the human body grow, divide and eventually die. In regular cells, the division process is organised and any mutations are removed via programmed cell death. In contrast, cancer cells do not follow the regular rules of cell division and cell death. The process is characterised by uncontrolled growth and replication, and the evasion of programmed cell death [7]. It begins with a mutation or gene alteration which affects cell homeostasis [8]. Malignant tumours invade surrounding tissue and have the capacity to spread to distant organs, also called metastases [9]. When this happens, the cancer is said to have metastasised. The term cancer encapsulates many different related diseases. So-called solid tumours are usually named based on the tissue or organ where the cancer originated.

Risk factors for cancer differ depending on the specific disease. Generally risk factors include age, alcohol intake, diet, obesity, radiation exposure and tobacco usage [10]. Such risk factors can also affect the prognosis of patients diagnosed with cancer. An additional important prognostic factor is stage at diagnosis. Stage refers to the size and location of the primary tumour, spread to the regional lymph nodes, and formation of distant metastases. Early stage refers to localised (and usually curable) disease, whereas advanced stage refers to distant metastatic disease.

3.1.2 Cancer in this thesis

The statistical methods in this thesis are applicable to most cancers provided certain assumptions hold. Two specific cancers that are analysed in this thesis are chronic myeloid leukaemia (Study II) and female breast cancer (Studies III and IV).

Chronic myeloid leukaemia

Chronic myeloid leukaemia (CML) is a haematological neoplasm which is characterised by the presence of the t9;22 translocation (the Philadelphia chromosome). CML is normally a slow growing leukaemia which can progress into a more advanced phase (either accelerated or

blastic). The majority of patients diagnosed with CML are diagnosed in the chronic phase and these patients currently tend to take treatment for life where the goal of such treatment is to prevent progression to a more advanced phase. The only cure for CML is to undergo allogeneic stem cell transplantation.

Treatment of CML has changed enormously over the years, which has resulted in a dramatic improvement in the survival of CML patients. Prior to the 1990s, CML patients experienced a median survival time of 3.2 years [11]. Allogeneic stem cell transplantation was the preferred treatment for those patients healthy enough to undergo it during the 1990s, and was shown to increase median survival time to 4.7 years [12]. A molecularly targeted anti-cancer agent called imatinib mesylate (IM), commonly known as Gleevec or Glivec, was introduced in 2001 in Sweden. This was after randomised controlled trials showed survival gains in comparison to other standard treatments [13]. Since then, several targeted therapies similar to IM have been introduced to treat patients who are resistant or intolerant to IM. Recently, results from clinical trials have suggested that IM can be discontinued for a large proportion of CML patients who experienced a very good response to treatment [14].

Risk factors for CML are largely unknown. However, exposure to radiation and increasing age have been shown to increase the risk of being diagnosed with CML. In Sweden, the median age at diagnosis is 60 years of age [15] and the incidence of CML is estimated to be 0.9 per 100,000 person-years; the incidence has largely remained constant over calendar year. Survival gains over calendar year have driven a dramatic increase in the prevalence of CML in Sweden, and prevalence is predicted to continue to increase over time [16]. Five-year relative survival is estimated to be 80% for Swedish patients diagnosed after the introduction of IM [11]. This estimate varies between different ages with better relative survival seen in younger patients.

Breast cancer

Breast cancer is the most common female cancer in Sweden. It represented 30% of all female cancers diagnosed in Sweden between 2011 and 2014. In 2014 there were 9,730 diagnoses of female breast cancer in Sweden [17]. The number of breast cancer diagnoses has been increasing over time whilst mortality has been decreasing [18, 19]. Risk factors for the incidence of breast cancer include family history of breast cancer, increasing age and breast density, among other factors [20].

Survival of breast cancer is generally high. The one-year age-standardised relative survival between 2011 and 2015 was estimated to be 0.98 (95% CI: 0.97-0.98) and at five years this was estimated as being 0.89 (95% CI: 0.89-0.90) [19]. There are many factors which affect survival of breast cancer patients including age, patient and tumour characteristics, and treatment. Mammographic screening was introduced in Sweden throughout the 1980s and 1990s and has been shown to increase the survival of breast cancer patients [21]. The Swedish National Board of Health and Welfare recommends that women aged 40-74 years be invited to mammographic screening every 18 to 24 months. Stage at diagnosis and socioeconomic status (SES) have been shown to affect survival of breast cancer patients. The effects of these factors are explored

in Study IV. There is evidence that patients diagnosed at a younger age often have a more aggressive disease and thus have a poorer prognosis [22] and it is known that advanced stage at diagnosis is also related to poorer prognosis. Additionally, there is evidence that survival differences between SES groups are partly due to differences in stage at diagnosis [23, 24, 25].

3.2 Survival analysis

3.2.1 A short overview of key concepts and measures in survival analysis

This thesis focuses on the survival of cancer patients and in order to describe this methodology, a short overview into the area of survival analysis is useful.

Survival analysis refers to statistical methods which analyse time-to-event data, where observations are followed over time until an event (or events) of interest. This thesis considers only one event of interest. The main components of this type of data are time, t_i, from a specified origin time, and an indicator for whether observation i experienced the event, d_i, or not. One of the elements of survival analysis that separates it from other statistical methods is censoring. Censoring occurs when the event status for an observation is not known [26]. There are different types of censoring, but in this thesis only right censoring is considered. Right censoring occurs when up until a certain time point no event has been observed. After this point we can no longer follow the observation; we know that sometime in the future the event could be observed but are no longer able to record it. This is the most common form of censoring in population-based cancer studies. It is important to note here that the assumption of non-informative censoring is made throughout this thesis, i.e., that censoring is independent of the event of interest. In addition, all survival data in this thesis are considered independent; when analysing correlated survival data other methods are required [27].

Three measures of importance which describe the relationship between the event time T and the event of interest are the survival function, the hazard function, and the cumulative hazard function. The survival function $S(t)$ can be defined as follows and is the probability the event of interest occurs after a certain time-point t,

$$S(t) = Pr(T > t).$$

The hazard function $h(t)$ defined below represents the instantaneous rate at time t of experiencing the event of interest and is presented per-person time

$$h(t) = \lim_{\delta \to 0} \frac{Pr(t \leq T \leq t + \delta | T \geq t)}{\delta}.$$

It is common to multiply the hazard rate by 1,000 or 100,000 and subsequently present the rates per 1,000 or 100,000 person-years. The cumulative hazard $H(t)$ is highly useful when it comes to modelling survival data, as described in section 5.1. The cumulative hazard function

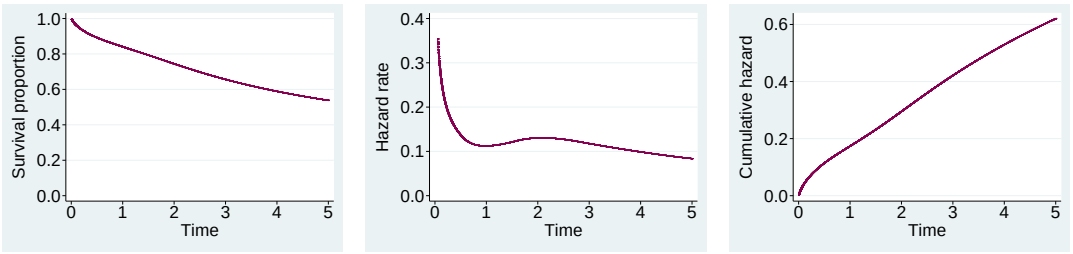
is defined as

$$H(t) = \int_0^t h(u)du.$$

Whilst the survival function is monotonic-decreasing and the cumulative hazard function is monotonic-increasing, the hazard function can be any non-negative function; see Figure 3.1 for an illustration of these measures. These three measures are all related to one another; the most important relationships are described below. Note that in this thesis mortality rate and hazard rate are used interchangeably to refer to the instantaneous rate of death within the cancer patient population.

$$h(t) = -\frac{d}{dt}\log S(t) \tag{3.1}$$

$$S(t) = \exp\left[-\int_0^t h(u)du\right] = \exp[-H(t)] \tag{3.2}$$



(a) **(b)** **(c)**

Figure 3.1: Examples of three important measures commonly used in survival analyses as functions of time. **(a)** shows the survival proportion, **(b)** the hazard function, and **(c)** the cumulative hazard function.

3.2.2 The Cox proportional hazards model

The Cox model [28] is the most common survival model used to date, and the second most cited statistical paper [29]. The Cox proportional-hazards model is formulated as follows

$$h(t;\mathbf{x}) = h_0(t)\exp(\beta\mathbf{x}) \tag{3.3}$$

where $\mathbf{x}$ represents covariates, β their estimated coefficients, and $h_0(t)$ represents the baseline hazard function, i.e., the hazard when all covariates are zero or at their specified baseline levels. In the Cox model, only the coefficients β are estimated, not the baseline hazard, thus hazard ratios (HRs) are estimated, but no absolute effects are directly estimated.

In contrast to the Cox model where the baseline hazard function is not directly estimated, we can make assumptions about the shape of the baseline function via parametric models; a brief description of these is given in section 5.1. One parametric model which is commonly used for the analysis of survival data is the Poisson model.

3.2.3 Modelling hazard rates using the Poisson model

Poisson models can also be used to model time-to-event data (Study III). Poisson models are commonly used for modelling count data but can also be used to model rates by introducing an offset into the model. An offset is a variable included in a model where no coefficient is estimated, i.e., the coefficient is forced to be equal to one. The log hazard rate can be modelled using Poisson models by introducing an offset of the person-time at risk, y, and using the relationship mortality $h(t) = n_d/y$, where n_d represents the number of deaths.

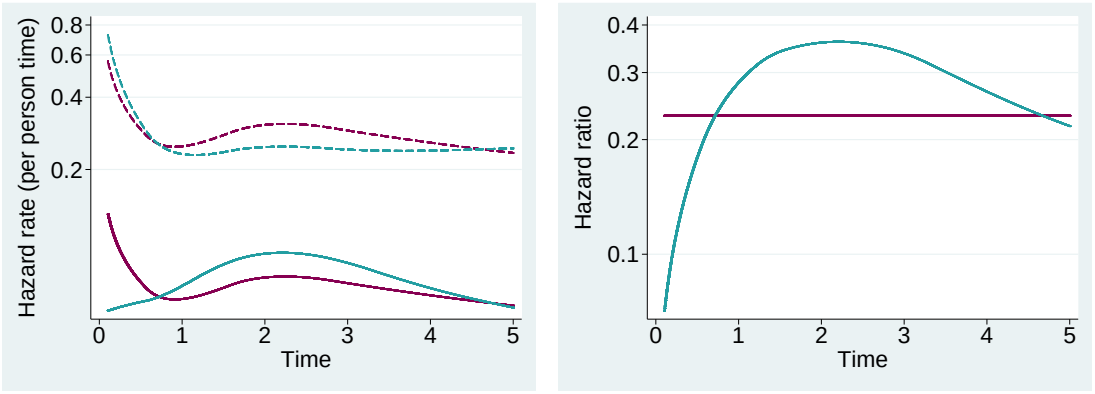
$$\begin{aligned} \ln(n_d) &= f(t) + \beta\mathbf{x} + \ln(y) \\ \Rightarrow \ln(n_d/y) &= f(t) + \beta\mathbf{x} \\ \Rightarrow \ln[h(t)] &= f(t) + \beta\mathbf{x} \end{aligned} \tag{3.4}$$

Note that the effect of time must be included in the model via $f(t)$. This is commonly done by splitting the time at risk into intervals, thus the number at risk and person-time at risk are also split into different time intervals. It is assumed that the hazard rate within each time interval is constant; these models are also called piecewise constant hazard models for this reason. It is also possible to subsequently smooth the piecewise constant hazards, for example by using splines [30]. As described above in equation (3.4), these models assume proportional hazards but this assumption can easily be relaxed by including interactions between variables and a function of time. A description of the proportional hazards assumption comes next.

3.2.4 Proportional versus non-proportional hazards

The term proportional hazards refers to the modelling assumption that the ratio of two hazard rates is constant over time, i.e., that the HR is constant over time. Allowing for non-proportional hazards means that two hazard rates are allowed to vary differently over time and thus the HR of these two rates can vary over time. Although it is often easier to interpret coefficients from a proportional hazards model, it is in fact common to see non-proportional hazards especially when follow-up is long (as is commonly the case for population-based studies).

In order to allow for non-proportional hazards, a model must include an interaction term between time and the covariate which is intended to have a time-dependent effect. Note that time-dependent effects or time-dependent coefficients are different from time-varying covariates; the latter refers to the situation where covariate values change over time, not necessarily their effects. A hazards model is described as having non-proportional hazards if one or more of the coefficients are allowed to vary over time. It is possible to extend the Cox model to incorporate non-proportional effects [31] although it is easier to estimate these effects when making some assumption about the shape of the underlying hazard, as is the case in parametric models.



(a) **(b)**

Figure 3.2: Proportional hazards versus non-proportional hazards. Purple lines are from a proportional hazards model whilst blue lines are from a non-proportional hazards model. **(a)** shows hazard rates for two levels of a covariate, solid and dashed lines represent the levels, and **(b)** shows the corresponding HRs.

4. Materials

The analyses and method development undertaken in this thesis were possible due to data contained in several population-based registers in Sweden. These population-based registers are used for many different purposes including incidence monitoring and assessment of quality of care and are also utilised within various research areas. In Sweden, every individual who is registered as a resident for at least a year is provided with a unique personal identification number. This number is used in many aspects of society including all contact with the Swedish healthcare system. The personal identification number is included in population registers and allows many of these registers to be linked. The linkage of registers creates a rich tool for research [32]. Note that data protection regulations require that the personal identification number is not available for research purposes, instead a pseudo-identification number is used. This thesis uses data from the Swedish Cancer Registry, the Breast Cancer Database Sweden (BCBaSe, a data merge based on data from the Regional Breast Cancer Clinical Quality Registers and other registers), and expected mortality data which are described in the following sections.

4.1 The Swedish Cancer Registry

The Swedish Cancer Registry (SCR) was established in 1958 and contains all individuals with a permanent residency in Sweden who have had a diagnosis of cancer. Six regional cancer centers were founded in the 1980s, where information on diagnoses of cancer are first sent to be registered, checked and coded. The national register is then compiled at the Swedish National Board of Health and Welfare (NBHW) who are responsible for the SCR [33]. The national register is updated once a year and approximately 50,000 malignant cases are recorded annually. By law, healthcare providers must report a diagnosis of cancer. In fact, most individuals diagnosed with cancer will be reported twice to the regional cancer centers; one report commonly comes from the initial diagnosis by a clinician and the second from a pathologist. Cancer diagnoses only recorded on death certificates are not recorded in the SCR. Information in the SCR includes patient characteristics such as sex and age, and medical data including site of tumour, date of diagnosis, and TNM stage (collected since 2004). Date of death and cause of death is obtained through linkage to the Swedish Cause of Death Register. The completeness of the register has been estimated to be above 96% [34]. Study II analyses Swedish CML patients using information from the SCR.

4.2 Breast Cancer Database Sweden (BCBaSe)

BCBaSe is a register merge based on information from three out of six Regional Breast Cancer Clinical Quality Registers (RQRBC). Since 1992, the six Swedish healthcare regions have reported to the separate RQRBC. BCBaSe contains information from the Stockholm-Gotland, Uppsala-Örebro, and North regions which together cover approximately 60% of the total population. Such clinical registers record detailed and individual-specific diagnostic and treatment information in the aim to evaluate and develop care for patients. Using data included in such registers for research purposes is becoming increasingly common.

BCBaSe contains information on clinical data including patient and tumour characteristics, clinical stage and treatment information from 1992 to 2012, alongside five controls per case matched on birth year and county of residence. BCBaSe was created by linking data from the RBCQR to data from the SCR, Cause of Death Register, and Total Population Register to confirm vital status, and for selecting controls. The data was further linked with the following data sources:

- The Longitudinal integration database for health insurance and labour market studies (LISA). This database contains, among others, information on SES from censuses, highest education achievement and disposable income for Swedish residents aged 16 years and over.
- The MiDAS database from the Swedish Social Insurance Agency. Information on sickness and other social benefits are available in BCBaSe from MiDAS.
- The National Patient Register. This register records all in-patient healthcare and specialist outpatient healthcare. The data linked to BCBaSe from this register can be used for quantifying comorbidities.
- The Multi-Generation Register. Information on relatives with breast cancer is obtained using this register in combination with the SCR. The Multi-Generation Register links individuals born in 1932 or later and who were registered in Sweden since 1961, to their parents.
- The Prescribed Drug Register. This records all prescriptions given to individuals in Sweden from 2005.

In Study III and IV, education information at the time of diagnosis from LISA is utlilised and used as a measure of SES for breast cancer cases (both Study III and Study IV) and controls (only Study III).

4.3 Expected mortality rates

In contrast to observed mortality rates among cancer patients, expected mortality rates are mortality rates calculated for a whole population. These rates are often used as a proxy for the

mortality diseased patients would have experienced had they been disease-free. Such rates are contained in tables referred to as life tables, or population mortality files. Expected mortality rates are commonly presented by age, sex and calendar year since these are often the only factors affecting mortality that are available for the whole population. Alternatively, methods can be implemented to additionally adjust these tables as illustrated in Study III of this thesis. The expected mortality rates used in this thesis are taken from the Human Mortality Database [35] which contains expected mortality rates for a number of countries including Sweden. The databases are available to download from the Human Mortality Database (`http://www.mortality.org/`) after becoming a registered user on their website. The Swedish expected mortality rates contained in the Human Mortality Database are obtained from Statistics Sweden. The use of expected mortality rates in population-based cancer survival analyses is described further in section 5.2.2 of this thesis.

5. Statistical methods and measures

5.1 Flexible parametric models

As described in the previous chapter, the most commonly implemented model when considering survival data is the Cox model. The Cox model makes no assumption on the shape of the baseline hazard function and is able to directly estimate relative effects, but not absolute effects. Although relative effects, like HRs, are very interesting, absolute effects help provide a greater understanding of the survival outlook. Alternatively to the Cox model, parametric models can be fitted. Parametric models make assumptions on the shape of the baseline hazard meaning they are able to directly estimate both relative and absolute effects. Flexible parametric models (FPMs) offer a good alternative to standard parametric models, such as the Weibull model, since they are able to capture simple and complex effects that such standard parametric models may struggle to capture [36]. Furthermore, due to their modelling of the baseline function, FPMs can be used to estimate many useful measures such as time-dependent HRs and hazard differences.

Royston and Parmar first introduced FPMs in 2001 as a tool for various forms of survival analysis [37]. These models have since been extended in a variety of settings including relative survival [38], to estimate cure [39] and incorporating random effects [40]. This thesis focuses on the use of FPMs in estimating relative survival and subsequently estimating the loss in expectation of life. The key to the flexibility of these models is their use of restricted cubic splines (RCSs). RCSs in FPMs model some transformation of the survival function. Here focus is on RCSs modelling the log cumulative hazard function and the log hazard function, although they can also model on other scales. RCSs are piecewise cubic polynomials joined together at points called knots [41]. The first and second derivatives at these join points are forced to be continuous to ensure the resulting function is smooth. The RCSs are additionally forced to be linear before the first, and after the last knot to ensure that data in the extremes do not affect the function excessively. Users select the number of knots, which dictate the degrees of freedom (the number of knots minus one), and determine how complex the resulting RCS is.

An RCS function, $s(\alpha; \gamma_0)$, with $\alpha = t$ or $\ln(t)$ and knots $k_1, ..., k_K$ is defined as

$$s(\alpha; \gamma_0) = \gamma_{00} + \sum_{l=1}^{K-1} \gamma_{0l} v_l(\alpha),$$

where the l^{th} basis function $v_l(\alpha)$ is defined as

$$v_l(\alpha) = \begin{cases} \alpha, & \text{for } l = 1 \\ (\alpha - k_l)_+^3 - \lambda_l(\alpha - k_1)_+^3 - (1-\lambda_l)(\alpha - k_K)_+^3, & \text{for } l = 2, ..., K-1, \end{cases} \tag{5.1}$$

and where k_1 and k_K are the boundary knots,

$$\lambda_l = \frac{k_K - k_l}{k_K - k_1} \qquad \text{and}$$

$$w_+ = \begin{cases} w & \text{if } w > 0 \\ 0 & \text{if } w \leq 0. \end{cases}$$

Splines are often orthogonalised to avoid any problems that can arise due to them being correlated. Figure 5.1 shows an example of a mortality rate modelled using an RCS with four degrees of freedom for malignant melanoma where the knot positions are illustrated with vertical lines. The scattered points show mortality rates estimated using the number of events and person time at risk in monthly intervals.

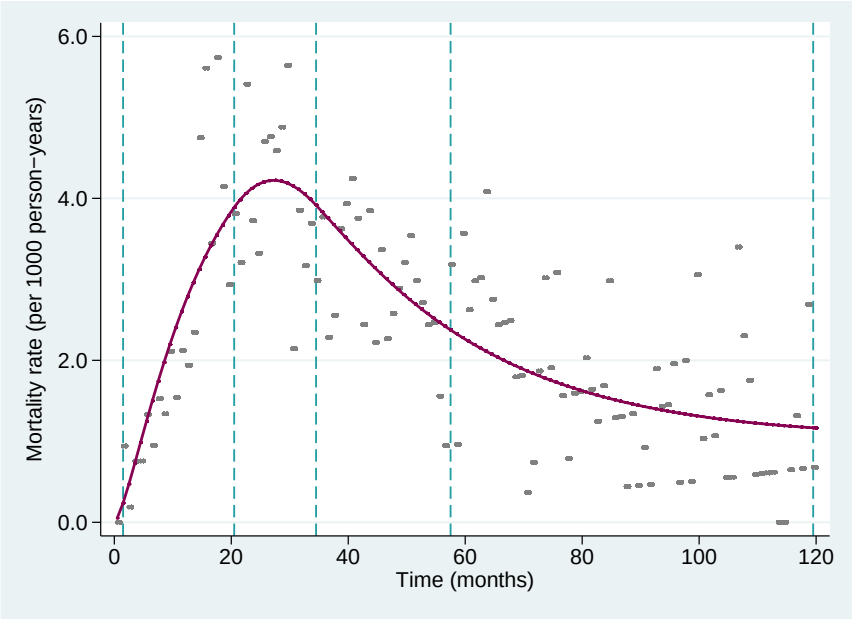


Figure 5.1: An illustration of an RCS used to model the log mortality rate for malignant melanoma patients. Points illustrate monthly mortality rates. The solid line represents the smooth RCS function with knots placed at the vertical dashed lines.

It is also important to note that similarly to the way RCSs are used in FPMs to model the baseline function, they can also be used to model continuous covariate effects. Instead of categorising continuous variables or making the assumption that the effect is, for example, linear, one can instead assume that the effect of a continuous covariate is captured by an RCS function with specified degrees of freedom. In all studies in this thesis, the effect of continuous covariates are modelled in this way.

5.1.1 Modelling on the log cumulative hazard scale

It is common for FPMs to be implemented on the log cumulative hazard scale. FPMs assuming proportional hazards on the log cumulative hazard, $\ln[H(t;\mathbf{x})]$, can be written as

$$\ln[H(t;\mathbf{x})] = s(\alpha;\gamma_{\mathbf{0}}) + \mathbf{x}\beta, \tag{5.2}$$

where t represents time, $\mathbf{x}$ are covariates, and $s(\alpha;\gamma_{\mathbf{0}})$ is an RCS function which here represents the baseline log cumulative hazard function. Non-proportionality is introduced into FPMs by including interactions between the covariates that are to have time-dependent effects, and time. It is common to include RCSs as a function of log time, but the untransformed time scale can be used and is useful when using age as the underlying timescale.

Time-dependent effects, or non-proportionality of J covariates, can be modelled by extending equation (5.2) in the following way

$$\ln[H(t;\mathbf{x})] = s(\alpha;\gamma_{\mathbf{0}}) + \mathbf{x}\beta + \sum_{j=1}^{J} s(\alpha;\gamma_{\mathbf{j}})\mathbf{x}_{\mathbf{j}}. \tag{5.3}$$

It has been shown that HRs estimated from FPMs are similar to those estimated from Cox models [36, 42]. Furthermore, Rutherford *et al.* showed that FPMs accurately capture both simple and complex hazard functions provided enough knots are specified for the spline function [43] under proportional hazards. In Study I of this thesis, we extended this work to assess FPMs under non-proportional hazards and similarly concluded that RCSs are able to capture a variety of hazard functions with small/minimal bias given an appropriate number of knots are specified.

Maximum likelihood estimation is used to fit FPMs where the log-likelihood contribution of the i^{th} individual is

$$\begin{aligned} \ln l_i &= d_i \ln[h(t_i)] + \ln[S(t_i)] \\ &= d_i \ln[h(t_i)] - H(t_i). \end{aligned} \tag{5.4}$$

Equation (5.4) is calculated using the relationship between the hazard function and the cumulative hazard function described in equation (3.2). This means that on the log cumulative hazard scale, the components of the likelihood function for fitting FPMs can all be analytically obtained. Equation (5.4) can be extended to include delayed entry (this is the case in Study IV where a period approach is used, see section 6.4.2) by adding the cumulative hazard at entry time:

$$\ln l_i = d_i \ln[h(t_i)] - H(t_i) + H(t_{0i}) \tag{5.5}$$

5.1.2 Modelling on the log hazard scale

One of the reasons for modelling on the log cumulative hazard scale rather than the log hazard scale is that numerical integration is required to model on the log hazard scale which can be computationally intensive. Moreover, the coefficients estimated from the cumulative hazards model can often be interpreted as log HRs; see appendix A.1 for an illustration of this. On the other hand, there are good reasons to accept the inconvenience involved in numerical integration and model on the log hazard scale. One such reason is that we can avoid problems that we encounter when modelling multiple time-dependent effects on the log cumulative hazard scale. The HRs of variables with time-dependent effects are dependent on other covariates with time-dependent effects when modelling multiple time-dependent effects on the log cumulative hazard scale; see appendix A.2 for an example of this issue.

An FPM on the log hazard scale, $\ln[h(t;\mathbf{x})]$ with J time-dependent effects can be defined as follows:

$$\ln[h(t;\mathbf{x})] = s(\alpha;\gamma_{\mathbf{0}}) + \mathbf{x}\beta + \sum_{j=1}^{J} s(\alpha,\gamma_{\mathbf{j}})\mathbf{x}_{\mathbf{j}}. \tag{5.6}$$

Modelling FPMs on the log hazard scale is easy to implement in Stata [44] using the `strcs` command; this is presented in a related publication [45] and described in Study I and Study III. The log-likelihood used when fitting FPMs on the log hazard scale is presented in equation (5.4) and can be rewritten in the following way:

$$\ln l_i = d_i \ln[h(t_i)] - \int_{t_{0i}}^{t_i} h(u)du. \tag{5.7}$$

The integral included in equation (5.7) cannot be calculated analytically due to the inclusion of RCSs in the hazard function. Thus, numerical integration methods must be implemented to calculate the log likelihood. Gaussian quadrature is a method of numerical integration that is implemented when modelling FPMs on the log hazard scale. Gaussian quadrature converts an integral into a summation in the following way:

$$\int_{-1}^{1} g(x)dx \approx \sum_{m=1}^{M} w_m g(x_m) \tag{5.8}$$

The summation is performed over pre-defined points called nodes, denoted m here, and weighted by w_m. In our application of Gaussian quadrature we calculate the integral between time t and the entry time t_0. Equation (5.8) can be altered to integrate between these time points in the following way [45, 46]:

$$\int_{t_0}^{t} g(x)dx \approx \frac{t-t_0}{2} \sum_{m=1}^{M} w_m g\left(\frac{t-t_0}{2}x_m + \frac{t_0+t}{2}\right) \tag{5.9}$$

Note that Gauss-Legendre quadrature is used in the Stata command `strcs` to numerically integrate the hazard function for each individual to get the cumulative hazard function needed

to maximise the likelihood function. Gauss-Legendre quadrature sets the weights $w_m = 1$.

5.1.3 Model selection

It is important to remember that there may be several good models which fit to analysis data. Still, model selection is highly important since we wish to draw accurate conclusions from the statistical method we choose. Selecting a reasonable model usually consists of 1) considering which type of model is appropriate for the data on hand, e.g., a Cox or FPM for survival data, and 2) selecting appropriate variables to include in the model which explain the variation in the outcome, e.g., categorical age, or an RCS for age, when modelling time-to-death.

An important aspect of model selection for FPMs is the selection of the degrees of freedom. In this context, the degrees of freedom can both refer to the complexity of the RCS that models the baseline function and separately the complexity of the RCS used to model the time-dependent effects. Information criterion are highly useful when selecting degrees of freedom for FPMs. The Akaike Information Criterion (AIC) [47] and the Bayesian Information Criterion (BIC) [48] are commonly used criteria that are calculated to aid FPM selection [36, 42]. Letting the likelihood be denoted L, p be the number of parameters within a model, and n be the number of observations, the AIC and the BIC are defined as follows:

$$\text{AIC} = -2\ln L + 2p \qquad \text{BIC} = -2\ln L + n\ln(p).$$

Note that within the time-to-event framework, n is often taken to be the number of events rather than the number of observations [49]. The models corresponding to the lowest AIC and BIC are the best-fitting models according to each criteria. It is important to be aware that the best-fitting model suggested by the AIC may not the same as the best-fitting model suggested by the BIC. Study I concluded that the AIC and BIC should be used to guide FPM selection.

5.2 Net survival

Net survival is defined as the survival probability of cancer patients when removing death due to other causes, i.e., survival if cancer patients could only die of their particular cancer. This may at first seem a strange measure to use since this is not what happens in reality. However, the advantage of this measure is that it is independent of differences in the non-cancer mortality, thus fair comparisons can be made between different populations and different time periods where the mortality rate due to causes other than cancer may vary. For example, consider a comparison between the survival of a certain cancer in two equal-sized populations. Population A are prone to have comorbidities and subsequently a high mortality rate due to causes other than cancer. In comparison to population A, population B has a lower mortality rate due to other causes. Suppose the mortality due to cancer in each of the populations is the same at all points in time. The all-cause survival in the two populations would indicate that population A has a worse survival than population B purely due to their increased risk of dying from

causes other than cancer. For this reason, it is interesting to estimate net survival, where we can remove mortality due to other causes and concentrate on cancer-specific survival.

There are two frameworks in which net survival can be estimated: 1) a cancer-specific framework, or 2) a relative survival framework. In this thesis we rely on estimating net survival via relative survival. A summary of terms and measures are given in table 5.1 to aid understanding and clarify the concepts presented in this chapter.

Survival measure	Details	Mortality measure
Net survival, $NS(t)$	Survival of cancer patients when removing other-cause mortality.	Net mortality
Relative survival, $R(t)$	Ratio of the all-cause survival and the expected survival. *Estimates net survival**	Excess mortality, $\lambda(t)$
Cause-specific survival	Survival of cancer patients where patients are followed until death due to cancer, other deaths are censored. *Estimates net survival**	Cause-specific mortality
All-cause survival, $S(t)$	Survival of cancer patients where patients are followed until death due to any cause. *Can also be called observed survival.*	All-cause mortality, $h(t)$
Expected survival, $S^*(t)$	Survival expected if the cancer patients had been cancer-free. Population survival is used as a proxy for this using population life tables.	Expected mortality, $h^*(t)$

*Under certain assumptions

Table 5.1: Summary of measures used in Chapter 5.

5.2.1 Relative survival framework

Within the relative survival framework there are several estimators of net survival, one of which is the relative survival estimator itself [50]. Relative survival, $R(t)$, is defined as the all-cause survival of the cancer patients, divided by the all-cause survival expected had these patients been cancer-free [51].

$$R(t) = \frac{S(t)}{S^*(t)} \tag{5.10}$$

The hazard analogue of relative survival is called the excess mortality. When the excess mortality equals zero, the survival in the cancer patient cohort and the expected survival are equal. Excess mortality, $\lambda(t)$, is the all-cause mortality observed in the cohort of cancer patients, $h(t)$, minus the expected mortality had the cancer patients been cancer-free, $h^*(t)$. It aims to capture mortality due to cancer.

$$\lambda(t) = h(t) - h^*(t) \tag{5.11}$$

Relative survival is a popular measure for estimating net survival since it does not require accurate cause of death information. Cause of death information can be unreliable [52], and even if reliable, it may not be clear if cancer is the primary cause of death. For example, cause of death data often exclude treatment-related deaths. In theory, such deaths should be recorded as due to cancer since these patients would not have died due to the treatment without the diagnosis of cancer. However, in practice, it is impossible to know whether a non-cancer death was related to cancer treatment or not.

Relative survival is often reported as an average measure for a particular population, commonly at a specific number of years post-diagnosis. This average measure is usually an internally age-standardised estimate of relative survival, i.e., a weighted average according to the age distribution found in the population under analysis. In contrast, external age-standardisaton is important when comparing populations whose age distributions may differ. When externally age-standardising, an alternative age distribution is forced on the population under study, for example, the age distributions described in the International Cancer Survival Standards [53].

In reality we expect the relative survival and excess mortality to vary by individual i. As a result, when we model we estimate relative survival on an individual level based on disease and patient characteristics. Individual level estimates are predicted in studies II, III and IV.

The relative survival estimator can be calculated from both a life table approach, or using a modelling approach. It is important to mention that there are several approaches in the life table framework, the Ederer II approach is the most commonly used method for estimating expected survival [54]. It has been shown that these estimators do not perfectly estimate net survival [55]. The correct estimate of net survival is:

$$NS(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{S_i(t)}{S_i^*(t)} = \frac{1}{n}\sum_{i=1}^{n}R_i(t). \tag{5.12}$$

However, the traditional life table approaches calculate this as the ratio of the average all-cause survival and the average expected survival:

$$R(t) = \frac{\frac{1}{n}\sum_{i=1}^{n}S_i(t)}{\frac{1}{n}\sum_{i=1}^{n}S_i^*(t)}. \tag{5.13}$$

Bias has been shown to be minimal when age-standardising relative survival estimates calculated as in equation (5.13), and that there may be some added benefits in terms of precision for longer term estimates of relative survival [56, 57]. In comparison to the traditional methods, Pohar Perme *et al.* proposed an unbiased estimator of marginal net survival. This Pohar Perme method directly estimates internally age-standardised net survival without stratifying by age, by accounting for informative censoring. Note that when we model relative survival we directly estimate $R_i(t)$ for each individual i, and can standardise as shown in equation (5.12).

In order for the marginal relative survival measure to estimate marginal net survival one must be willing to make two assumptions. Firstly, that the cancer population and the general population are independent conditional on the matching variables, most commonly age, sex

and calendar year, and variables included in the analysis. Secondly, that the mortality in the general population, or the expected mortality, represents the mortality the cancer patients would have experienced had they been cancer-free.

There are situations when estimating net survival is not of interest, rather we are interested in estimating survival in the real world where patients can die of other causes as well as their cancer. Competing risks methods that estimate these crude probabilities are not discussed in this thesis. Instead we focus on a measure called the loss in expectation of life (LEL), a real-world survival measure which is discussed in section 5.3.

Flexible parametric relative survival models

As described previously, this thesis focuses on estimation of relative survival via modelling. FPMs can be implemented to estimate relative survival by utilising information from the expected mortality [42]. These models work in a very similar way to the FPM as presented in equation (5.2) but instead model the log cumulative excess hazard, $\Lambda(t)$:

$$\ln[\Lambda(t;\mathbf{x})] = s(\alpha;\gamma_{\mathbf{0}}) + \mathbf{x}\beta. \tag{5.14}$$

The cumulative excess hazard can be calculated in a similar way to the excess hazard shown in equation (5.11):

$$\Lambda_i(t) = H_i(t) - H_i^*(t). \tag{5.15}$$

FPMs which model the log cumulative excess hazard can easily be extended to allow for non-proportionality by including interactions in a similar way to those shown in equation (5.3). Similarly to when modelling the log cumulative hazard using FPMs, maximum likelihood estimation is used when using FPMs to estimate relative survival. The likelihood function is written as follows [58]:

$$\ln l_i = d_i \ln[h^*(t_i) + \lambda(t_i)] + \ln[S^*(t_i)] + \ln[R(t_i)].$$

Since $S^*(t_i)$ does not depend on any model parameters, the likelihood function can be simplified in the following way:

$$\ln l_i = d_i \ln[h^*(t_i) + \lambda(t_i)] + \ln[R(t_i)]. \tag{5.16}$$

This means that the likelihood can be maximised by merging in the expected mortality rates at the time of death without the need for numerical integration or time-splitting.

5.2.2 The expected survival/mortality

The expected survival is defined as the survival that would be expected had the cancer patients not been diagnosed with cancer. Survival probabilities and mortality rates are often obtained from nationwide population life tables [35, 59] and are matched to characteristics of the cancer

patients. Such expected survival probabilities or mortality rates are used when calculating relative survival.

In order to interpret relative survival as net survival, we make two assumptions as previously stated, 1) that the all-cause survival and the expected survival are conditionally independent given certain variables, usually age, sex and calendar year, and 2) that the expected survival is representative of what the cancer patients would have experienced had they not have had cancer. In practice, there are many variables other than age, sex and calendar year that both the all-cause survival and the expected survival differ by, for example the presence of comorbidities or SES. However, this is usually only problematic if one wishes to quantify relative survival by one such variable, but does not have expected mortality rates adjusted by that variable [60, 61], thus violating assumption 2. For example, suppose one is interested in estimating relative survival by SES where SES-specific life tables are not available. In truth the expected mortality is higher for the low SES, and lower for the high SES than the expected rates available. If relative survival is estimated by SES using life tables without consideration of SES, the relative survival for the low SES group will be an underestimate, and the relative survival for the high SES will be an overestimate.

In Study III we present methodology which can utilise information contained in control populations to further adjust expected mortality rates. Previous work in creating adjusted life tables includes using data available for the whole population [62, 63, 64], modelling grouped data [65] or using available mortality and prevalence information [62, 66, 67]. Alternatively, Howlader *et al.* suggest using a cause-specific approach instead of a relative survival approach if appropriate life tables are not available [68].

One other consideration when using expected survival from population life tables for the estimation of relative survival, is that the cancer patients under study are commonly included in the general population used to estimate the expected mortality. This would imply that the expected mortality is slightly higher than it should be if the cancer patients had been excluded, but in practice, this inclusion of cancer patients makes little difference to relative survival estimates of specific cancers [69].

5.3 The loss in expectation of life

Relative survival as an estimate for net survival has great advantages for comparing cancer patient survival between different groups. The interpretation of relative survival as an estimator of cancer patient survival if cancer patients could only die of their cancer is, however, difficult to grasp since it is a hypothetical situation. In contrast, such 'real-world' measures as the LEL can be estimated to compliment estimates of relative survival, or to answer a different question about the impact of a cancer diagnosis in the real-world.

The LEL gives a measure of the life years lost to disease, or the reduction in the life expectancy due to disease; in this thesis we focus on the effects of a diagnosis of cancer. In comparison to net survival, the LEL can be offered as an intuitive measure which quantifies survival over patients' whole life span. The simple interpretation of LEL could be useful for

non-statisticians understanding of the effect of a disease and the probable prognosis of diseased patients. The interpretation remains simple even when complex modelling is required to capture underlying patterns in survival data. It can also be used to both quantify the burden of cancer in society and on average for an individual.

The LEL is calculated as the difference between the life expectancy in a cancer-free population and that in a cohort of cancer patients matched by age, sex and year; again the general population is used as a proxy for this former value. Life expectancy, also known as mean survival time, is calculated as the area under the survival curve:

$$\text{Life expectancy} = \int_0^\infty S(t)dt, \tag{5.17}$$

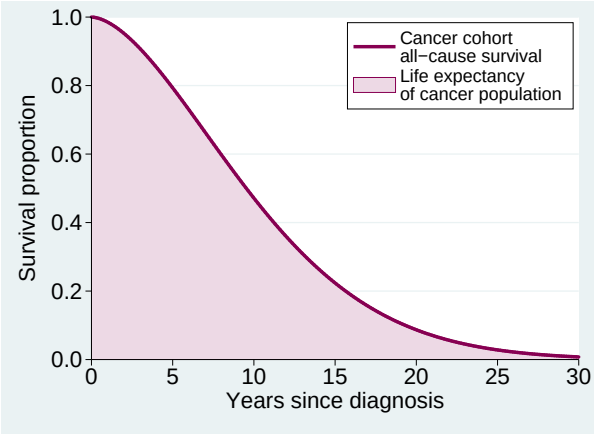
where the survival curve continues until time reaches infinity. Letting $\mathbf{x}$ denote covariates, of which $\mathbf{x}'$ are the covariates also in the population life table and are a subset of $\mathbf{x}$, the LEL can be written as:

$$LEL(\mathbf{x}) = \int_0^{t_{max}} S^*(t, \mathbf{x}')dt - \int_0^{t_{max}} S(t, \mathbf{x})dt. \tag{5.18}$$

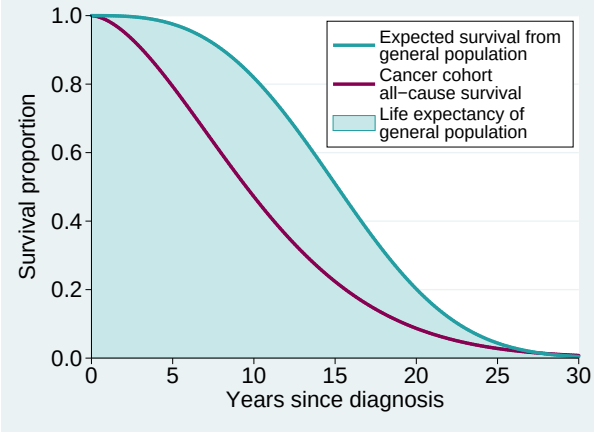
$S(t, \mathbf{x})$ is the all-cause survival in the cohort of cancer patients, and $S^*(t, \mathbf{x}')$ is the expected survival, or, the all-cause survival in the general population. The two integrals in this equation calculate the life expectancy in the general population and the life expectancy in the cohort of cancer patients, thus quantifying the change in the life expectancy associated with a diagnosis of cancer. Instead of integrating to infinity as shown in equation (5.17) for the life expectancy, the LEL contains integrals between zero and a specified time t_{max} where both the survival functions $S(t, \mathbf{x})$ and $S^*(t, \mathbf{x}')$ are assumed to be zero. Calculation of the LEL is illustrated in Figure 5.2.

5.3.1 Extrapolation of survival curves

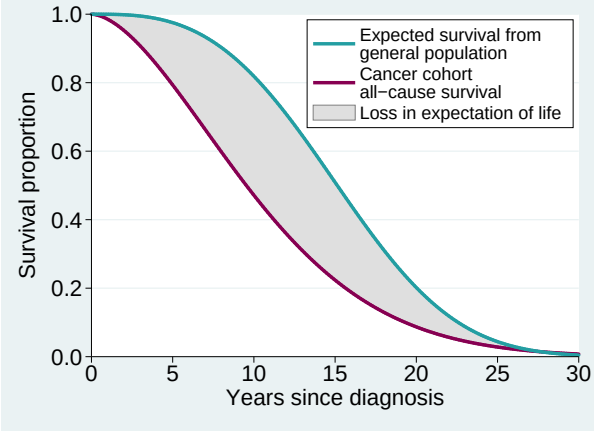
In order to calculate the integrals in equation (5.18), we require the population to have been followed for long enough so that the survival has reached zero. This can be problematic since follow-up is often too short to observe all deaths, thus, extrapolation of the survival curve is undertaken. We can make reasonable assumptions about the expected survival extrapolation, either using predictions for the future, or by assuming that the expected survival in the future years will be the same as that in the most recently observed year. Extrapolation of the survival within the smaller and more specific cohort of cancer patients is more problematic. One method of extrapolating the all-cause survival of the cancer cohort is to fit a parametric model to the observed all-cause survival based on the data that is available and extrapolating that curve. However, even if this model fits well to the observed data, it may extrapolate poorly [70]. Hakama and Hakulinen suggested extrapolating the relative survival and then combining with expected survival instead by using the relationship described in equation (5.10) [71]. In order to extrapolate using relative survival, the LEL equation in (5.18) can be re-written in the



(a)



(b)



(c)

Figure 5.2: Illustration of how the LEL is estimated. **(a)** shows the all-cause survival observed in the cancer patients; the area under the curve represents the life expectancy, or mean survival, of the cancer cohort. **(b)** shows the expected survival taken from the general population and the life expectancy of this group as the area under the curve. **(c)** shows the LEL as the difference between the two life expectancies shown in **(a)** and **(b)**.

following way:

$$LEL(\mathbf{x}) = \int_0^{t_{max}} S^*(t, \mathbf{x}')dt - \int_0^{t_{max}} S^*(t, \mathbf{x}') \cdot R(t, \mathbf{x})dt, \tag{5.19}$$

and can extrapolate relative survival for the estimation of LEL in one of the following three ways:

- **Assumption 1: Assume that the excess hazard is zero after a certain time point**. This is the same as assuming a point of statistical cure. This is implemented by fitting a flexible parametric cure model [72]. The excess hazard is zero after the point of cure, resulting in a constant cumulative excess hazard after the point of cure. Thus, the FPM used to extrapolate relative survival when using the statistical cure assumption, imposes this constraint that the cumulative excess hazard is a constant after the last knot in the model.

- **Assumption 2: Assume that there is a non-zero, constant excess hazard after a certain time-point.** In this FPM, the cumulative excess hazard can increase after the last knot, but is constrained to increase in a constant way.

- **Assumption 3: Assume that the log cumulative excess hazard follows the linear trend predicted by the model.** The use of RCSs in FPMs means that the function is linear after the last knot; this assumption extends this linear function into the future.

Hakama and Hakulinen suggested, outside of the FPM methodology, that the extrapolation of relative survival be implemented by assuming either the first or second assumption listed [71]. Andersson *et al.* subsequently suggested using an FPM-based approach to model the log excess cumulative hazard using these two extrapolation methods and additionally via the linear trend predicted by the model [70]. Andersson *et al.* also showed that the FPM approach is accurate for a variety of cancers for the assumption of cure, provided this is a reasonable assumption for the particular cancer of interest, and when extrapolating based on the linear trend estimated from the model. The extrapolation of relative survival works better than the extrapolation of all-cause survival since, for most cancers, the excess mortality is highest early on, and is low after a number of years. Thus, after a number of years, the extrapolation is driven by the expected survival extrapolation which is easier by making reasonable assumptions. Note that extrapolation via this method may not be advisable if the excess mortality is not low and decreasing towards the end of the available follow-up. This is also illustrated by Figure 5.3 which shows an example of the extrapolation.

The LEL as described by Andersson *et al.* has been implemented in studies involving breast cancer patients in the UK [73], for Swedish AML patients [74], among invasive cancer patients in Australia [75], and in colon cancer patients [76], in addition to those discussed within this doctoral thesis. Studies II and IV implement the approach by Andersson *et al.* where the linear trend is extrapolated.

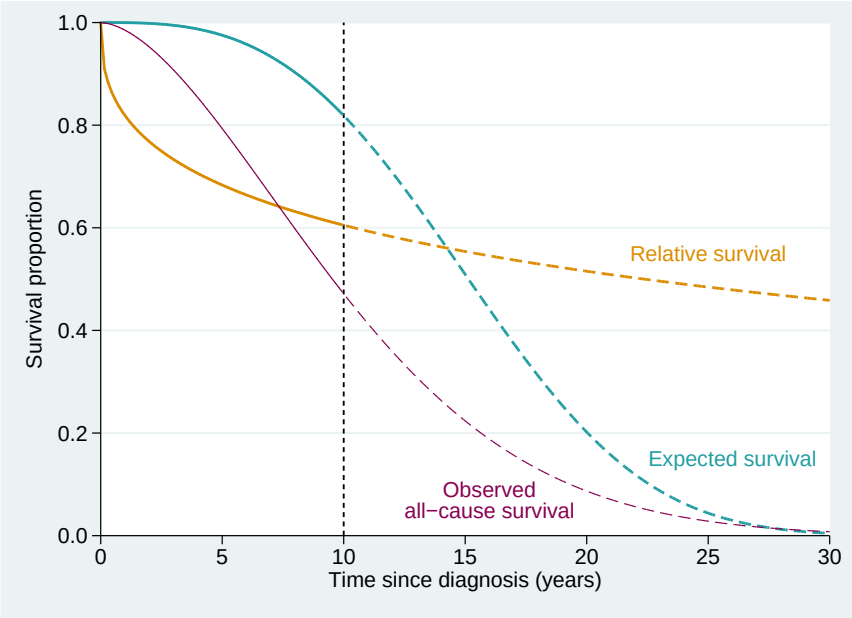


Figure 5.3: Example of relative survival extrapolation compared to observed all-cause survival extrapolation. The shape of the relative survival curve is easier to capture due to low excess mortality after a number of years in comparison to the observed all-cause survival.

The LEL is presented here as a cancer-specific measure. It is possible to quantify the reduction in the life expectancy for other diseases, but the assumptions made using the methodology in the cancer setting to estimate the LEL may not be reasonable in other diseases. Estimating the LEL in the way described in this thesis relies on the assumption that the cancer patients would experience the survival seen in the general population if they had been cancer-free. Using the general population in this way has been shown to not be problematic for most cancers but may not be reasonable when considering other diseases. For example, this assumption may not be appropriate for patients with coronary heart disease since they have a higher prevalence than the general population of other comorbidities and risk factors which could result in death due to causes other than coronary heart disease [77].

5.3.2 Measures related to the loss in expectation of life

There are several methods which are closely related to the LEL, or that aim to estimate a similar quantity [78]. Measures which estimate the impact a diagnosis has on the potential life span include the years of life lost (YLL). This is a measure which is closely related to the LEL. The YLL estimates the impact a diagnosis of cancer has via the number of life years lost due to cancer. This can be calculated in several different ways. Burnet *et al.* describe the YLL as the difference between the age at cancer death and the life expectancy for someone of the same age [79]. Equivalently this can be calculated as the number of deaths multiplied by the life expectancy at the age of death [80, 81]. One problem with calculating the YLL in this way is that it will never be zero since the life expectancy is calculated conditional on surviving until the age of interest. Mettlin *et al.* describe an alternative way to calculate the YLL in which the

age at death is subtracted from a set cut off age, where 65 years has commonly been used as the cut off [82, 83]. Any deaths after the cut-off age contribute nothing to the YLL. Selecting a relevant cut-off age for this method is arguably subjective. Although the LEL and the YLL are similar in that they aim to quantify premature mortality and both account for changes in both the cancer-mortality and the mortality in the general population, there are some additional weaknesses of the YLL [84]. Specifically, YLL relies on correct cause of death information which would be problematic if death information is not reliable. It also disregards any treatment-related deaths. However, even if cause of death is accurate, the measure cannot be used for a specific cohort since it can only be calculated for those patients who have died. This also means it is not optimal for considering the potential prognosis of patients [83]. On the other hand, one strength of the YLL is that it can be used for diseases other than cancer since it is not calculated based on the same assumptions as the LEL.

Other measures that are included in this thesis in Study IV are the proportion of expected life years lost (PELL), total life years lost, gain in life years, and the postponable deaths. A generalised summary of these measures and the LEL are presented in Table 5.2. The PELL is calculated as the LEL divided by the life expectancy that would have been expected had the patient not been diagnosed with cancer. Whilst the LEL is highly age dependent since younger ages have more remaining life years to potentially lose, the PELL is a measure which is more comparable over different ages [85]. The total life years lost is the LEL multiplied up, e.g., by the number of cases diagnosed in a calendar year. This quantifies the loss on a population level and provides a measure of cancer burden on the whole population. Both the gain in life years and the postponable deaths require a counterfactual situation that can be compared to the actual observed data. For example, in Study IV we compare the LEL estimated from the observed data, to the LEL that would have been estimated had the relative survival been the same in all education groups. In this way we can determine the potential effects of eliminating differences in education group between the number of life years lost (estimated by the LEL), or the number of deaths (postponable deaths). The gain in life years can also be multiplied up to provide a population-level measure. Note that postponable deaths are often also called avoidable deaths. The term postponable was chosen to indicate that the measure is highly time-dependent and that no death is truly avoided if followed up for a long enough time.

One interesting observation by Rutherford *et al.* was that the postponable deaths measure and the LEL are related [86]: the gain in life years can be calculated by integrating the postponable deaths as so:

$$\begin{aligned}\text{Gain in life years} &= \left[\int_0^{t_{max}} S^*(u)du - \int_0^{t_{max}} R(u)\cdot S^*(u)du\right] \\ &\quad - \left[\int_0^{t_{max}} S^*(u)du - \int_0^{t_{max}} R_{alt}(u)\cdot S^*(u)du\right] \\ &= \int_0^{t_{max}} R(u)\cdot S^*(u)du - \int_0^{t_{max}} R_{alt}(u)\cdot S^*(u)du\end{aligned}$$

$$= \int_0^{t_{max}} \underbrace{S^*(u)\,[R(u) - R_{alt}(u)]}_{= \text{postponable deaths}}\, du. \tag{5.20}$$

Measure	Quantifies	Equation
LEL	Life years lost or loss in life expectancy	see section 5.3
PYLL	Proportion of remaining life years lost due to the diagnosis of cancer	$\text{PYLL} = \frac{\text{LEL}}{\text{Expected remaining life years}}$
Total LYs lost (TLYL)	Total life years lost in population	$\text{TLYL} = N \cdot \text{LEL}$
Gain in LYs (GLY)	Gain in LYs if an alternative situation was experienced by patients	$\text{GLY} = N \cdot (\text{LEL}_{alt} - \text{LEL})$
Postponable deaths (PD)	Postponable deaths if an alternative situation was experienced by patients	$PD = N \cdot \{[R_{alt}(t) - R(t)] \cdot S^*(t)\}$

LEL_{alt} and RS_{alt} refer to the LEL and relative survival, respectively, that would be estimated under an alternative or counterfactual situation. N is the number of individuals at risk.

Table 5.2: Description of measures related to the LEL that are estimated in Study IV.

6. Overview of the studies

6.1 Study I

Study I of this thesis is titled *Capturing Simple and Complex Time-Dependent Effects Using Flexible Parametric Survival Models: A Simulation Study* and focuses on 1) assessing the ability of RCSs in FPMs in capturing time-dependent effects, 2) comparing this ability when modelling the log hazard and the log cumulative hazard, and 3) determining whether model selection for FPMs should be based on the AIC or the BIC.

6.1.1 Motivation

FPMs are becoming more popular for modelling survival data as an alternative to the Cox model and other parametric models. This is due to their ability to model the baseline function in a flexible way. Although these models have been assessed under the proportional hazards assumption [43], it is common for this assumption to be violated. Users of FPMs often rely on information criterion, usually the AIC and/or the BIC to select the optimal number of degrees of freedom, but it is not clear if one method outperforms the other.

6.1.2 Simulation strategy

We simulated time-to-event data under nine scenarios with differing complexities. These nine scenarios were a combination of three different baseline hazard functions and three different time-dependent functions. The scenarios were denoted 1A-1C, 2A-2C, 3A-3C where the baseline hazard used was denoted by the numbers 1-3 and the letter referred to the time-dependent effect. Note that the paper includes scenarios 1-3 which are scenarios 1A, 2B, and 3C, respectively. Baseline functions were taken from Rutherford *e*t al. Time-dependent effects are shown in Figure 6.1 where the underlying HR is a function of time. The survival times were simulated using methods discussed in Bender *et al.* [87] and subsequently Crowther *et al.* [88]. Hazard functions for both levels of the binary covariate x are shown in Figure 6.2.

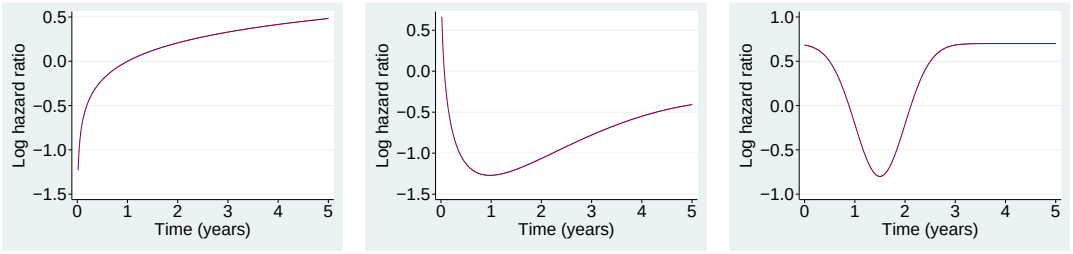


Figure 6.1: Time-dependent effects included in the simulation scenarios. **(a)** shows the time-dependent effect in scenario **A** simulated using a real time-dependent effect of deprivation status found in breast cancer patients [89]. **(b)** shows the time-dependent effect used in scenario **B** which is based upon an effect seen for stage in colon carcinoma patients [90]. **(c)** shows the time-dependent effect used for scenario **C** based on a fictional effect [91].

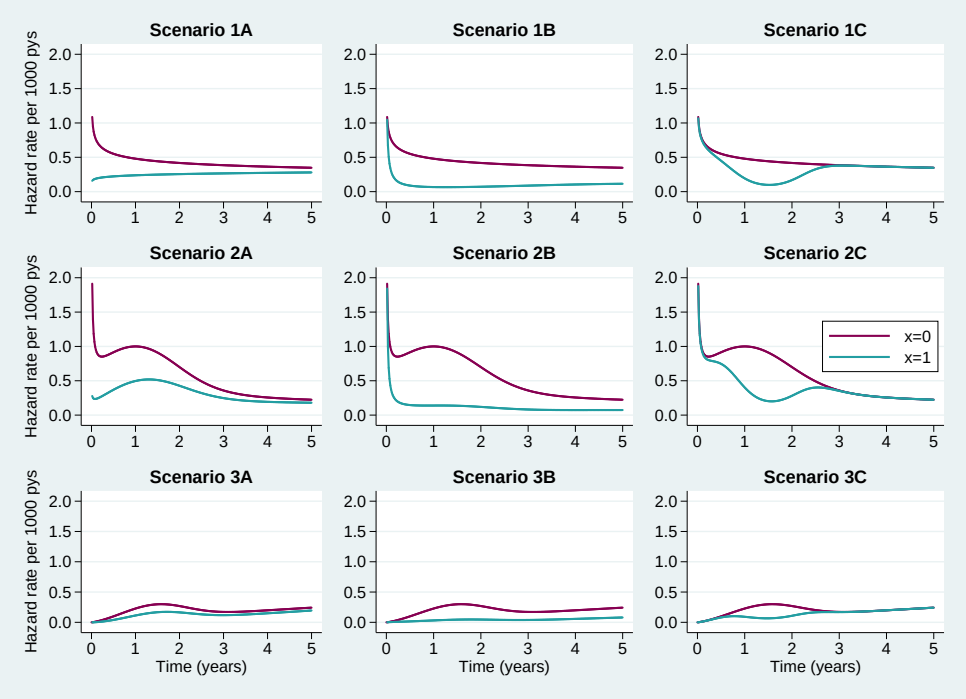


Figure 6.2: The hazard rates from nine simulated scenarios.

FPMs were fitted to the simulated data sets with varying degrees of freedom. Survival proportions, hazard rates and HRs were predicted from these models and the AIC and BIC calculated; see Table 6.1 for a summarised description of the simulation strategy. We assessed the performance of each model by comparing the predictions to the true values and calculated the bias, root mean square error, and coverage for each degrees of freedom selected. Bias is defined as the difference between the expected value of an estimator and its true value whereas root mean square error reflects the spread of the estimates. Coverage measures the proportion of the total number of simulations where the confidence interval contains the true value.

We wished to utilise the fact that we knew the truth in this simulation setting by also assessing how appropriate model selection by the AIC and BIC was. To do this we calculated a measure called the Symmetrised Kullback-Leibler distance (SKL). The fitted model with the

Number of simulations	500
Sample size	• 1,000 • 10,000 (only for models on the log cumulative hazard scale)
Models fitted	FPMs with all combinations of 3,5,7 degrees of freedom for the baseline & 1,3,5,7 degrees of freedom for the time-dependent effect on the: • log cumulative hazard scale • log hazard scale
Estimates of interest	• log HR at 1, 3, and 5 years • hazard rates at 1, 3, and 5 years • survival proportions at 1, 3, and 5 years

Table 6.1: Summary of the simulation strategy.

smallest SKL is the model closest to the truth. In the context of hazard estimation, the SKL can be written as [92]

$$\text{SKL}(\hat{h},h) = \frac{1}{n}\sum_{m=1}^{n}\int_{T_{0m}}^{T_m} \ln\left[\frac{h(t,X_m)}{\hat{h}(t,X_m)}\right]\left[h(t,X_m)-\hat{h}(t,X_m)\right]dt, \tag{6.1}$$

where t is time, T_{0m} and T_m are the entry and exit times respectively, and X_m represents the covariates for each observation $m = 1,2,...,n$. The SKL measures the difference between the fitted model and the true model [93, 94, 95]. We compared the optimal model, i.e., the optimal degrees of freedom, chosen by the AIC and the BIC to that selected by the SKL in order to determine which of the two information criteria were more accurate. Note that the SKL can only be calculated in a setting where the underlying truth is known.

6.1.3 Main results

Outside of a simulation setting, it is common that the degrees of freedom chosen when modelling using an FPM would be based on the values of the AIC and/or the BIC. Thus, Table 6.2 and Table 6.3 summarise the main results from this study by presenting bias from the models that were chosen by the AIC and the BIC. Note that this means that the optimal degrees of freedom chosen by the AIC could be different for different simulated datasets; similarly for the degrees of freedom chosen by the BIC. Tables 6.2 and 6.3 show the absolute bias of predictions estimated from models on the log cumulative hazard scale for 1,000 and 10,000 observations, respectively. The results show that the bias overall was low for all scenarios when using the AIC- and BIC- selected model. Bias increased, but was still small, when considering the log HRs in comparison to the survival proportions. At the same time, bias in the log HR was lower when using 10,000 observations in comparison to using 1,000 observations. For example, the bias in the log HR for scenario 3C at time $t = 5$ was 0.089 when using the AIC for 1,000 observations. When calculating this based on 10,000 observations, the bias reduced to 0.041. Bias in the hazard rates and log HRs also generally increased with time. It is important to remember that follow-up ended at five years, thus it may be reasonable for the models to perform less well at

five years.

		AIC model, absolute bias					BIC model, absolute bias				
	t	$S_0(t)$	$S_1(t)$	$h_0(t)$	$h_1(t)$	logHR	$S_0(t)$	$S_1(t)$	$h_0(t)$	$h_1(t)$	logHR
	1	0.000	0.000	0.005	0.002	0.004	0.002	0.000	0.004	0.002	0.003
1A	3	0.000	0.000	0.002	0.001	0.003	0.001	0.001	0.002	0.001	0.003
	5	0.000	0.000	0.006	0.003	0.007	0.000	0.000	0.006	0.004	0.005
	1	0.001	0.002	0.003	0.006	0.007	0.004	0.004	0.023	0.016	0.020
2B	3	0.002	0.002	0.029	0.006	0.003	0.005	0.005	0.047	0.004	0.012
	5	0.000	0.000	0.002	0.011	0.027	0.002	0.000	0.040	0.014	0.085
	1	0.000	0.000	0.001	0.001	0.010	0.008	0.010	0.006	0.000	0.002
3C	3	0.000	0.000	0.003	0.002	0.021	0.004	0.001	0.030	0.012	0.192
	5	0.000	0.001	0.011	0.002	0.089	0.000	0.000	0.032	0.001	0.184

Table 6.2: Summary of the absolute bias in the survival proportions, hazard rates and log HRs for models on the log cumulative hazard scale using degrees of freedom selected by the AIC and BIC; based on 1,000 observations. Here the notation $S_0(t)$ refers to the survival at time t estimated for $x = 0$. Results are presented to three decimal places.

		AIC model, absolute bias					BIC model, absolute bias				
	t	$S_0(t)$	$S_1(t)$	$h_0(t)$	$h_1(t)$	logHR	$S_0(t)$	$S_1(t)$	$h_0(t)$	$h_1(t)$	logHR
	1	0.001	0.000	0.000	0.000	0.000	0.002	0.001	0.001	0.001	0.000
1A	3	0.000	0.000	0.002	0.000	0.002	0.000	0.001	0.002	0.000	0.002
	5	0.001	0.000	0.004	0.000	0.005	0.001	0.000	0.004	0.002	0.004
	1	0.000	0.000	0.019	0.002	0.001	0.002	0.001	0.015	0.004	0.002
2B	3	0.000	0.001	0.018	0.007	0.008	0.001	0.002	0.021	0.007	0.006
	5	0.000	0.000	0.018	0.010	0.017	0.000	0.001	0.012	0.011	0.029
	1	0.001	0.001	0.003	0.001	0.001	0.000	0.000	0.007	0.004	0.003
3C	3	0.000	0.000	0.000	0.002	0.010	0.000	0.000	0.001	0.003	0.008
	5	0.000	0.000	0.009	0.001	0.041	0.000	0.000	0.010	0.001	0.044

Table 6.3: Summary of the absolute bias in the survival proportions, hazard rates and log HRs for models on the log cumulative hazard scale using degrees of freedom selected by the AIC and BIC; based on 10,000 observations. Here the notation $S_0(t)$ refers to the survival at time t estimated for $x = 0$. Results are presented to three decimal places.

6.1.4 Log hazard versus log cumulative hazard

One thing we were particularly interested in was whether modelling on the log cumulative hazard scale outperformed modelling on the log hazard scale. Due to the numerical integration, models on the log hazard scale took longer to fit and there were more convergence issues on this scale. For these reasons the analyses for 10,000 observations were restricted to modelling on the log cumulative hazard scale. Since this study, more work has been put into optimising the numerical integration in the `strcs` command in Stata which has improved both the speed and convergence, but there do still remain disadvantages to modelling on the log hazard scale in comparison to the log cumulative hazard scale. In terms of bias, slightly larger biases were present on the log hazard scale in comparison to the log cumulative hazard scale. However, these remained small when considering the AIC- and BIC- selected models. Moreover, it is important to remember that more degrees of freedom will be required to model a hazard

function in comparison to a cumulative hazard function. The results in this study did not necessarily indicate that modelling on the log cumulative hazard scale should be preferred, which was reassuring when considering our next methodological developments were planned for FPMs which model the log hazard function.

6.1.5 Discussion

The results from this simulation study indicate that FPMs accurately capture time-dependent effects, but that some thought into the selection of the degrees of freedom is required in order to avoid biased results, especially if one is interested in hazard rates or HRs. It is also important to remember that some scenarios included in this study were chosen purposely to be extreme, i.e., the time-dependent effect C. Guidance for users in selecting the degrees of freedom is important so we aimed to provide this via the section *Advice for users of flexible parametric survival models* in the paper. However, extensions of these recommendations may be advantageous to users, especially new users. Such advice is highly important when the user decides what complexity of model they will fit to their data.

In the same manner, FPMs have been criticised for 1) the need to select the degrees of freedom, and 2) the need to select the knot positions. Although we did not examine knot positions in this study, these are commonly placed at evenly distributed percentiles of the event times [42] and it has been shown that unless extreme choices are made, changing the knot positions rarely makes a difference to the model estimates [36, 42]. This conclusion is similar to our findings considering the selection of the degrees of freedom, i.e., that unless extreme values in the degrees of freedom are chosen, changing the number of knots makes little difference to the model estimates, particularly the survival proportions. As an alternative to making these selections, one could use penalised methods which do not require the user to select the knot positions and the number of knots. However penalised methods still rely on the user selecting a maximum degrees of freedom [96]. Comparisons between FPMs using the AIC to select the degrees of freedom and penalised methods show that both work well in capturing effects from data with proportional hazards [96].

On the other hand, the selection of knots does not need to be a weakness of these methods if appropriate model selection is undertaken. Selection of degrees of freedom being guided by the AIC and/or the BIC can be seen as a positive aspect of these models. Sensitivity analyses are always required to ensure that the conclusions are not sensitive to the model selected and by performing FPM analyses as suggested, fitting several models and comparing the information criterion to help select a good model, sensitivity analyses are built into the model-selection process. Seeing that the results and conclusions do not change based on a reasonable choice of the degrees of freedom is a positive thing that users should want to state.

There remain several disadvantages associated with this simulation study. One of the main disadvantages is the selection of the variable with the time-dependent effect. In our simulation study, we created a binary variable with a time-dependent effect. This limits the implications of the results since real-data analyses may 1) have more than one variable with a time-dependent

effect, and 2) have non-binary variables with a time-dependent effect. Performing a simulation with more time-dependent effects of other types of variables is a potential extension of this work, but we believe that this study remains essential for implementations of FPMs in data where proportional hazards is not a reasonable assumption.

6.1.6 Conclusion

Generally, the results indicated that FPMs capture simple and complex effects under non-proportional hazards and that the selection of degrees of freedom should be guided by both the AIC and the BIC. Modelling on both the log cumulative hazard and log hazard scales performed equally well. Users should remember to use both information criteria to help guide their selection and remember that many models will fit well. It is also advisable to ensure that the model selected captures patterns, in say the hazard function, that seem reasonable for the subject area. The results from this study offer reassurance to users of FPMs who are using these as tools for modelling complex data which often has non-proportionality. Although further work into multiple time-dependent effects with differing complexities is desirable, this work is a necessary assessment of FPMs which also offers an insight into the process of fitting these models for current and potential new users.

6.2 Study II

Study II titled *Life Expectancy of Patients With Chronic Myeloid Leukemia Approaches the Life Expectancy of the General Population* is an application of the LEL to Swedish CML patients.

6.2.1 Motivation

CML is a slow growing leukaemia which can become fatal if not treated. Consequently, most patients currently take CML treatments over their whole lifetime. A previous study by Björkholm *et al.* [11] showed that the relative survival of CML patients diagnosed after 2001 in Sweden was greater in comparison to previous calendar years, and concluded that this improvement was mainly due to the introduction of the treatment IM in 2001. Since CML is a chronic disease, it is highly interesting to see how these changes in treatments translate into changes in the life expectancy via the LEL.

6.2.2 Materials and methods

The data used for this analysis comprised of 2,662 Swedish CML patients diagnosed between 1973 and 2013. This data was obtained from the SCR. Patients were followed from their date of CML diagnosis until their date of death, emigration date, or end of follow-up (31 December 2013); whichever came first. We included patients who were aged 50 and above at diagnosis. This was because extrapolation of relative survival in patients younger than 50 years at diagnosis has not been studied [70].

FPMs were fitted with five degrees of freedom to model the baseline log excess cumulative hazard. Main effects of age at diagnosis (RCS with four degrees of freedom), year of diagnosis (RCS with four degrees of freedom) and sex were included in the model. Interactions between year of diagnosis and age at diagnosis, sex and year of diagnosis, and sex and age at diagnosis were modelled. Time-dependent effects of age at diagnosis, year of diagnosis and sex were also included. Only the linear RCS term of continuous variables were included in interactions and time-dependent effects. The extrapolated relative survival was estimated using the linear predictor from this model (assumption 3 described in section 5.3.1) and was then multiplied by the expected survival to get the observed survival. The expected survival was obtained from the mortality in the Swedish population available until 2012, and extrapolated by Statistics Sweden. We used an FPM to model relative survival and subsequently predicted the LEL and the proportion of expected life lost from this model.

6.2.3 Main results

The main results are shown in Figure 6.3, where the remaining life years, or remaining life expectancies, of CML patients and the general population are presented over year of diagnosis, by age and sex. This figure shows that there have been vast improvements in the life expectancy over calendar period, with improvements beginning in early 1990s for younger

patients, and improvements beginning mid- to late-1990s for the older patients. These improvements increase for all patients until they reach almost the same life expectancy as the general population in 2013, with predictions indicating that patients diagnosed with CML in 2013 will lose approximately three of their remaining life years lost due to their diagnosis of CML.

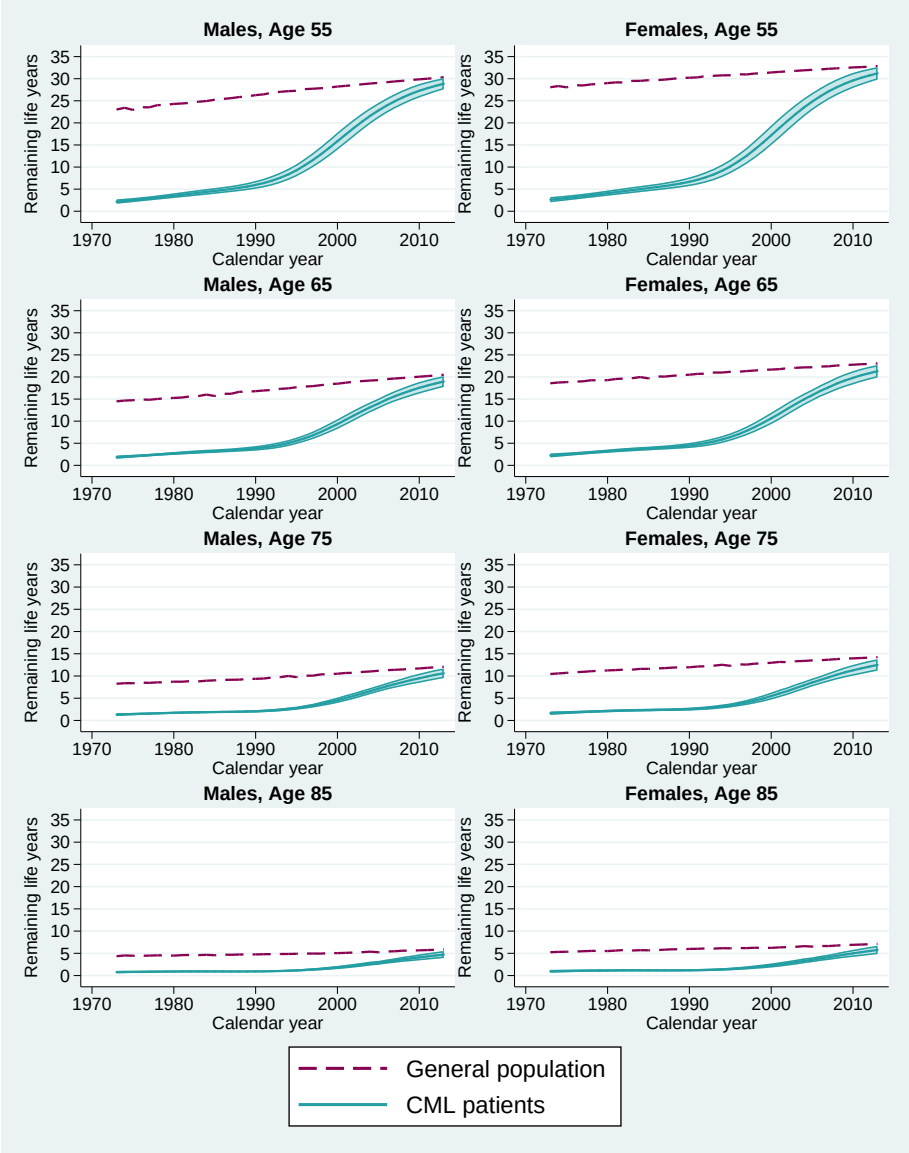


Figure 6.3: The remaining life expectancy of the general population and Swedish CML patients over year of diagnosis by sex and age at diagnosis.

6.2.4 Sensitivity analyses

Modelling assumptions can often be criticised without relevant sensitivity analyses. One sensitivity analysis that we undertook was to determine the effect that changing the degrees of freedom had on the results. We altered the degrees of freedom used to model both the baseline log cumulative hazard and the time-dependent effects. Figure 6.4 shows the effect of this on the LEL for patients diagnosed with CML over calendar year. This figure shows that for those aged 75 and 85 years at diagnosis, the results were insensitive to changes in the degrees of freedom. For those aged 55 and 65 years, changing the degrees of freedom made little difference to the life years lost up until approximately year 2000. After year 2000, changing the degrees of freedom made a small difference to the life years lost. These small differences seemed to be driven by changes in the degrees of freedom chosen to model the time-dependent effects rather than the baseline. The conclusions drawn in the paper are insensitive to these changes with perhaps the exception of the results seen for those aged 65 using one degree of freedom to model the time-dependent effect. The model with one degree of freedom for the time-dependent effect suggested that the LEL begins to level off for those aged 65 years. Both the AIC and the BIC values indicated that the best fitting models required more than one degree of freedom for the time-dependent effect and thus, this is likely an effect of undermodelling.

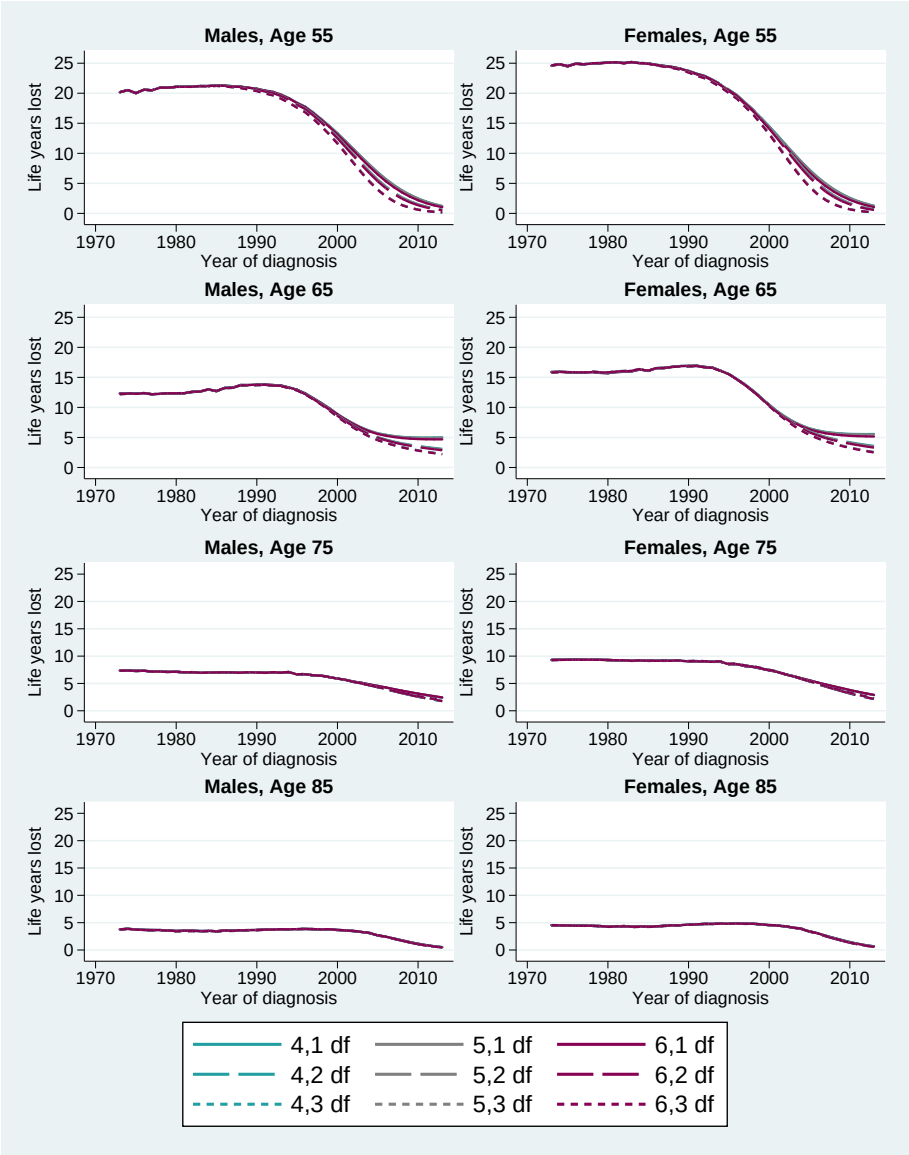


Figure 6.4: LEL for males and females diagnosed with CML over calendar year. Results are shown for models with differing degrees of freedom (df) where the first number refers to the degrees of freedom used to model the baseline log cumulative hazard and the second number refers to the degrees of freedom used to model time-dependent effects.

6.2.5 Discussion

Previous studies in this area have quantified survival of CML patients using relative survival. Björkholm *et al.* concluded that 'major improvement' was seen during the period 2001-2008 mainly due to to the use of IM [11]. Although both our study and the study undertaken by Björkholm *et al.* saw improvement from the early 1990s, we did not observe such a vast improvement in the life expectancy or LEL from 2001 when IM was introduced. It is important to remember that relative survival and LEL measure different aspects of cancer patient survival, and these differences could have contributed to these inconsistencies. Furthermore, it is possible that this difference could have been affected by how calendar year was modelled in the two studies. We modelled calendar year continuously using an RCS, whereas Björkholm *et al.* categorised calendar year; categorising can hide effects which would be seen otherwise when using a continuous variable. We additionally calculated the relative survival to extend the analyses undertaken by Björkholm *et al.* with follow-up until 2013; see Figure 6.5 for the five-year relative survival over time since diagnosis. Improvements prior to 2001 are seen when modelling year continuously which illustrates how modelling selections can make differences. Nevertheless, the introduction of IM had a large part in the improvements shown in the study by Björkholm *et al.* and Study II, and modelling calendar year continuously or via categories would not have altered these conclusions.

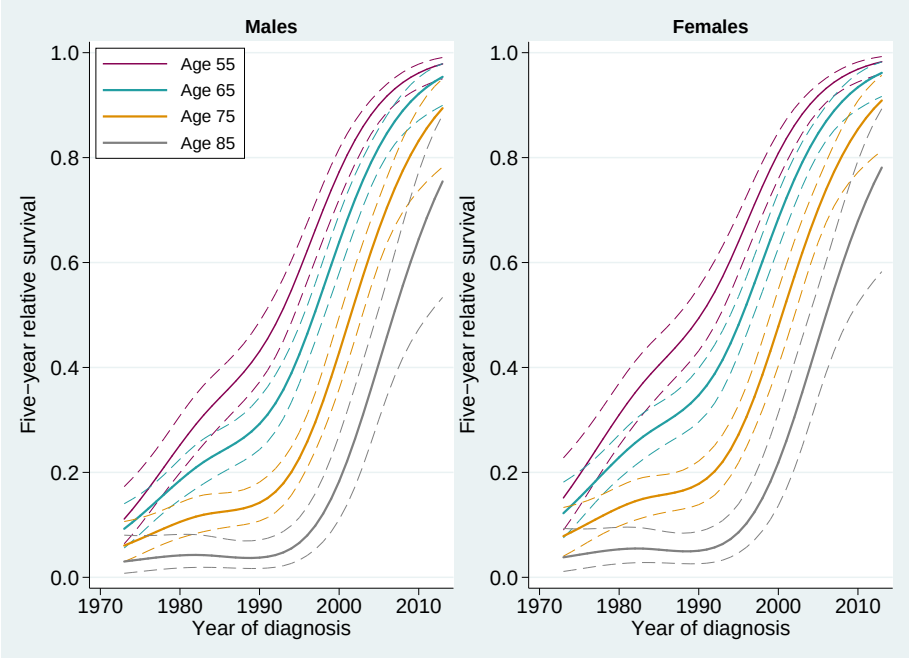


Figure 6.5: Five-year relative survival ratio over calendar year by age and sex.

Our study shows great improvements over time for CML patients in Sweden, but this does not mean that work in treating these patients is over. More research is needed to assess potential negative late effects of IM; lack of follow-up in our study means assessing this was not possible. Such problems seem plausible since there have been studies which have found an increased risk of cancer [97] and cardivascular deaths [98] involved in such treatments. This means that

the life expectancy of CML patients may never actually reach that of the general population. There also remains a subset of patients who are diagnosed in the acute phase who have an inferior response to targeted treatments such as IM [99]. It is also important to consider that the results presented in Study II may not be applicable outside of Sweden since individuals in other countries may not have the same sort of access to healthcare [100].

Another implication of these results is the impact on the Swedish healthcare system. Although incidence of CML has remained stable over the years, the survival gains have resulted in an increase in the prevalence of CML in Sweden. Although this could suggest a large economic increase for the Swedish healthcare system, Ohm *et al.* found that IM would pass cost-effectiveness levels accepted by healthcare authorities. Furthermore, there has been a recent focus on the possibility of discontinuation of targeted therapies like IM in CML patients. Clninical trials in this area show that it is possible to safely discontinue IM in well-responding CML patients [14, 101].

6.2.6 Conclusion

This study illustrates how life expectancy and LEL can be presented to better understand cancer prognosis in a population. CML patients diagnosed in the most recent calendar years are expected to lose, on average, around three years of their remaining life due to their diagnosis of CML, a vast improvement in comparison to previous calendar years which is largely due to treatment advances.

6.3 Study III

Study III presents methodology for adjusting expected mortality rates by additional variables that are available within a control population. The publication is titled *Adjusting Expected Mortality Rates Using Information From a Control Population: An Example Using Socioeconomic Status*.

6.3.1 Motivation

Expected mortality rates are often used as a comparitive rate for diseased patients. To accurately estimate measures such as relative survival, LEL, and standardised mortality ratios, expected mortality rates must be representative of the diseased population had they not been diseased. Analyses may be limited to quantifying these measures by age, sex, calendar year, and disease-specific factors since expected mortality rates from the population are often only presented for the former three factors. If analyses are performed by other factors which the expected mortality rates are additionally expected to vary by, then bias may occur. Although information additional to age, sex and calendar year are often unavailable on a population level, control populations that represent a random subset of the general population may be available with such data. We provide methods to adjust expected mortality rates by additional factors by utilising information contained in an available control population.

6.3.2 Developed methodology

Using SES information in the form of highest education achieved (low, medium, high) from 333,212 controls in BCBaSe, we adjusted Swedish population mortality rates using two different methods, 1) a Poisson approach, and 2) an FPM approach. Both approaches aimed to estimate the difference in mortality rates between the whole population of Sweden, h^*, and the control population in BCBaSe, h_c, and to quantify these differences by SES. In this way, one can estimate a value (adjustment factor, ρ) that the population mortality rates must be multiplied by to get SES-adjusted rates.

Equation (6.2) illustrates how the SES-specific adjustment factors ρ_j for different SES groups j, can be estimated as a function of attained age a and attained year y using the expected rates in the control population and the general population.

$$\frac{h_c(a,y,j)}{h^*(a,y)} = \exp\left[\sum_{j=1}^{3} \rho_j(a,y) \cdot \text{SES}_j\right] \tag{6.2}$$

Rearranging this equation illustrates how we estimated the adjustment factors in the Poisson and the FPM by modelling the rates in the control population, and including an offset of the

mortality rate taken from the general population:

$$\ln[h_c(a,y,j)] = \sum_{j=1}^{3} \rho_j(a,y) \cdot \text{SES}_j + \ln[h^*(a,y)]. \tag{6.3}$$

Note that there is no constant term, or baseline coefficient, included in the model and that all SES indicators are included with coefficients. This is so that we can directly estimate the effect of each SES group on the difference between the two mortality rates, i.e., directly estimate the adjustment factors.

Poisson model approach

We split on both the attained age and the calendar year timescale for the Poisson approach and modelled the rates in the control population in a similar way to the Poisson model presented in equation (3.4) by modelling the number of deaths, and including an offset for the person-time at risk. Consequently the Poisson model approach included two offsets. The first offset contained the person time at risk in order to allow modelling of rates rather than counts, and the second contained the expected mortality rate found in the general population in order to estimate the adjustment factors.

Flexible parametric model approach

The FPM approach applied in this study was undertaken using the `strcs` command I developed in Stata [45] which models on the log hazard scale. In order to estimate an adjustment factor for each of the SES groups, and not model the baseline log hazard, we first smoothed the expected mortality rates taken from the general population, and constrained the baseline log hazard spline function in the FPM. This worked in the following way:

1. Smooth the expected mortality rates in the population life table as a function of attained age and attained year (here $h_s^*(a,y)$ represents the smoothed expected mortality rate found in a population life table):

$$\ln[h_s^*(a,y)] = s_a(a;\gamma_a) + s_y(y;\gamma_y). \tag{6.4}$$

2. Model the difference between the expected mortality rate and the mortality rate in the control population (in order to estimate adjustment factors) using an FPM:

$$\ln[h_c(a,y,j)] = \sum_{j=1}^{3} \rho_j(a,y) \cdot \text{SES}_j + \ln[h_s^*(a,y)] \tag{6.5}$$

$$= \sum_{j=1}^{3} \rho_j(a,y) \cdot \text{SES}_j + s_a(a;\gamma_a) + s_y(y;\gamma_y) \tag{6.6}$$

where the third term in equation (6.6), $s_a(a;\gamma_a)$, is the baseline spline function described

in equation (5.6) but is constrained to be the the spline function in the smoothed rates model (6.4). The fourth term, $s_y(y;\gamma_y)$, is also forced to be the same as the spline in equation (6.4). Applying constraints in this way allows the FPM to estimate adjustment factors as required. Note that in general this means that at least one of the continuous variables in equation (6.4) must be modelled using RCSs.

Relative survival of Swedish breast cancer patients

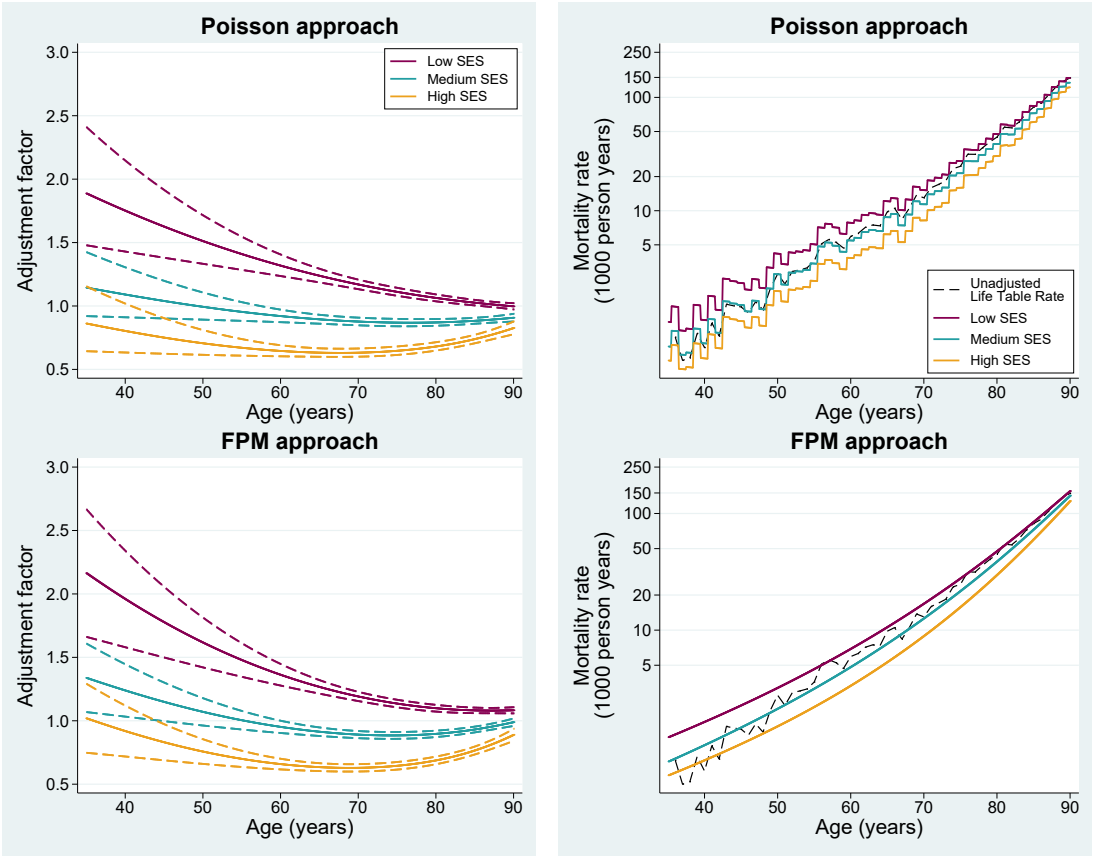
We used the two described methods to adjust expected mortality rates by SES. We then used these SES-adjusted expected mortality rates, plus the unadjusted rates, to estimate relative survival of female breast cancer patients recorded in BCBaSe. Consequently, FPMs were fitted to model the log excess cumulative hazard function using 1) original unadjusted expected mortality rates, 2) SES-adjusted expected mortality rates estimated from the Poisson approach, and 3) the SES-adjusted expected mortality rates estimated from the FPM approach.

Uncertainty of estimates using adjusted expected mortality rates

It is common to assume that mortality rates contained in population life tables carry no uncertainty since they are calculated from the whole population. It has been shown that making this assumption makes little difference to the variance of model-based relative survival estimates [102]. However, there is additional uncertainty to account for when using adjusted life tables (using the previously described methodology) in subsequent analyses. In this study we implemented a parametric bootstrap method in combination with Rubin's Rules [103] to account for this additional uncertainty. We used these methods to estimate confidence intervals for relative survival estimates of breast cancer patients contained in BCBaSe. We also selected 10% and 1% random samples of the control population to determine the impact of smaller control populations on the uncertainty in the relative survival estimates.

6.3.3 Main results

The main results from this study showed that both methods estimated similar adjustment factors; larger differences were seen for ages below 50 years. These differences were not seen in the SES-adjusted mortality rates since the mortality rates in those under 50 years was low; see Figure 6.6. As expected, differences between the relative survival estimated using SES-adjusted and non-adjusted expected mortality rates were seen, with the relative survival estimates using unadjusted expected mortality rates overestimating the differences between SES groups. Confidence intervals for five-year relative survival using the bootstrapping method showed increased confidence intervals when using 1% of the control population ($\approx$ 3000 individuals).



(a) **(b)**

Figure 6.6: (a) shows adjustment factors estimated from the Poisson approach, and the FPM approach. Adjusted rates for the two approaches are shown in **(b)**.

6.3.4 Sensitivity analyses

Several different sensitivity analyses were undertaken to determine the effect of the model chosen to estimate the adjustment factors on 1) the resulting adjustment factors and 2) the SES-adjusted mortality rates. These sensitivity analyses included changing the way the age variable was modelled, changing degrees of freedom for the splines included in the models, and investigating the effect of including interactions. We also considered how the estimates altered when changing the model used for smoothing the unadjusted expected mortality rates. Minor changes were seen in the SES-adjusted expected mortality rates when small changes were made to both the models specified for 1) smoothing the unadjusted expected mortality rates and 2) estimating adjustment factors. The adjustment factors were more sensitive to model changes than the SES-adjusted rates. Most larger differences in the adjustment factors were seen in younger individuals where fewer events occur, thus resulting in little difference on the mortality scale.

One additional sensitivity analysis that we undertook was to determine both the effect of the size of the time-splitting interval, and the effect of smoothing in the Poisson models. In the paper we present two methods, but there are other 'hybrid' models that could be considered.

Figure 6.7 illustrates these along with Table 6.4. The models presented in this figure and table are Poisson models which either split time more finely, or use smoothed expected mortality rates. Similar to previous findings, the adjustment factors on the left of Figure 6.7 were more sensitive to these changes whereas the mortality rates on the right were fairly insensitive.

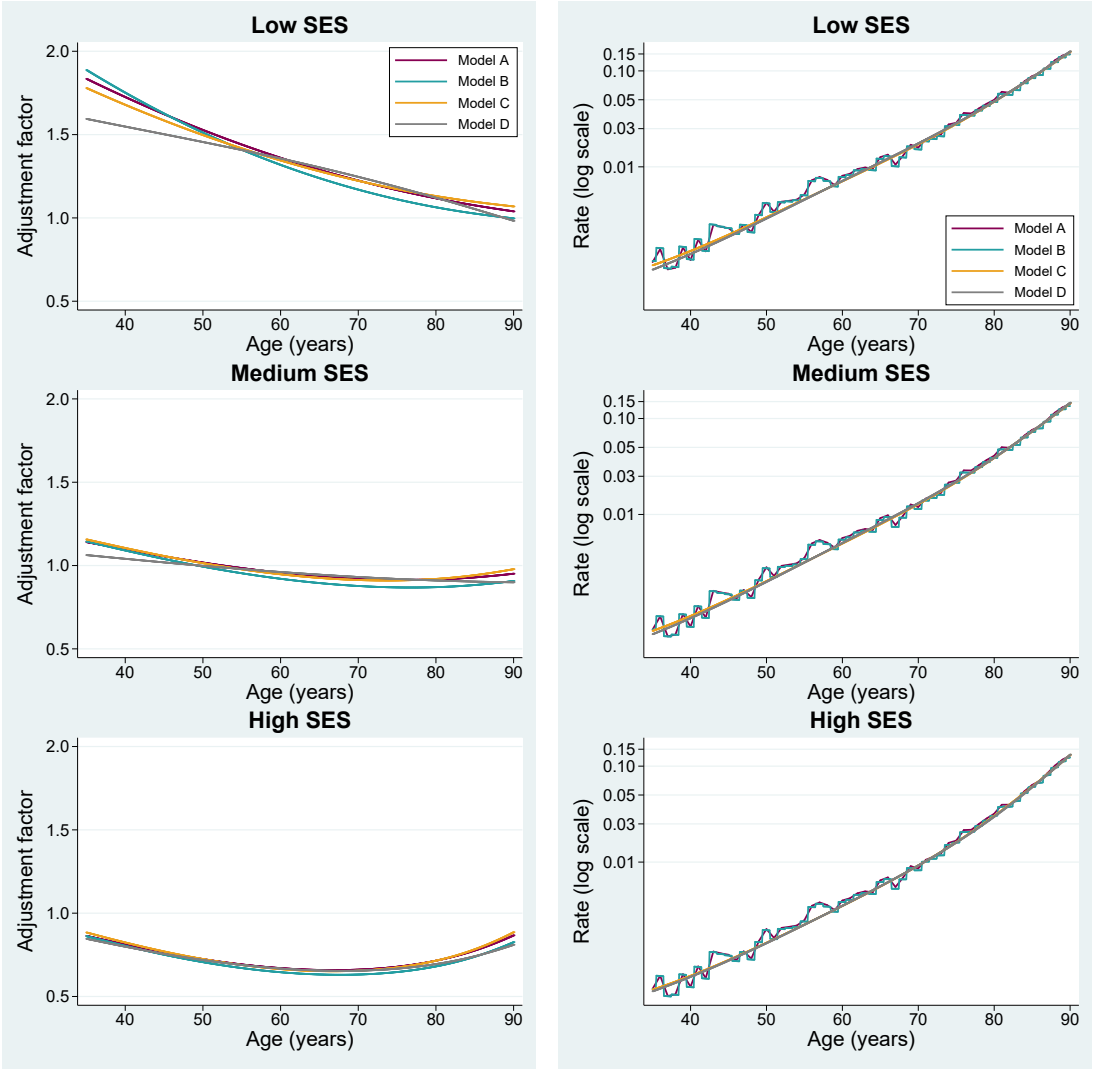


Figure 6.7: (a) shows adjustment factors and **(b)** shows SES-adjusted mortality rates from sensitivity analyses of the Poisson approach. Models are explained in Table 6.4.

Model	Control population timescales $h_c(a, y, j)$	General population timescales $h^*(a, y)$
A	Split yearly	Split yearly
B	Split monthly	Split yearly
C	Split monthly	Smoothed rates
D	Continuous	Smoothed rates

Table 6.4: Description of each of the Poisson models in the sensitivity analysis presented in Figure 6.7. Note that model **B** is the Poisson model presented in the paper and that model **D** is the Poisson equivalent of the FPM model, i.e., parameters in the adjustment factor model are constrained in order to include the smoothed expected mortaltity rates.

6.3.5 Discussion

Potential users of these methods may wonder which of the two methods to use. We suggest that individuals who may not be familiar with FPMs use the Poisson approach since this is the simpler method to apply. Alternatively, if one is worried about computational speed, it is advisable to use the FPM approach. As suggested in the sensitivity analyses presented here, it is possible to use the Poisson model in a similar way to the FPM model by smoothing the expected mortality rates and constraining estimates in the Poisson model. This method is a good alternative for individuals who are not familiar with FPMs and are worried about the computational speed involved in time-splitting. When the methods were being developed, the Stata command `strcs` [45], did not allow for the inclusion of an interaction between SES and year which could have easily been incorporated into the Poisson model.

One of the advantages to using the methodology described is that we utilise information both from the control population and the general population. Using the information from the general population reduces the uncertainty of our subsequent estimates of interest, for example, five-year relative survival, in comparison to only using information from the control population. Additionally, it is sensible to assume that the (SES-) adjusted expected mortality rates have a similar shape to the shape found in the general population; we make this assumption when utilising information from both the control population and the general population.

One of the greatest advantages to this methodology is that, contrary to the control population used in the publication, the control population need not be related to the diseased population of interest. By control population we refer to a random subset of the general population whose information can be utilised. Control populations are used in many aspects of research. These control populations could be utilised via these methods provided the control population is representative of the general population of interest. The size of such control populations should also be considered since very small control populations may not accurately represent the general population and will increase uncertainty in subsequent analyses.

6.3.6 Conclusion

These methods offer the opportunity to further estimate survival measures by non-standard variables. One only requires a control population that is representative of the general population,

i.e., a random subset, that need not be matched to the diseased population. Analyses can be limited by the lack of detailed factors on a population level and this work offers the possibility for further research which quantifies estimates such as relative survival in terms of factors, for example, SES, comorbidity and ethnicity.

6.4 Study IV

Study IV titled *Quantifying the Potential Gain in Life Years for Breast Cancer Patients in Sweden if Differences Between Education Groups Could be Eliminated* evaluated the impact that differences in education level could have on the number of deaths and life years lost for Swedish breast cancer patients. In this study, we attempted to answer such questions as 'How many life years could be saved if differences in stage at diagnosis between education groups could be removed?', and 'How many deaths would we expect to avoid in a calendar year if differences in breast cancer survival were removed between education groups?'.

6.4.1 Motivation and materials

It is known that survival differences exist between breast cancer patients with different SES, and stage has been shown to play a role in these differences [24]. Measuring the survival outcomes that could be gained if differences in SES were removed provides a greater understanding of the impact of these differences. We aimed to quantify the impact of differences between education groups in terms of the stage-distribution at diagnosis and stage-specific breast cancer survival.

62,121 breast cancer patients recorded in BCBaSe were used in the analyses. Highest educational achievement data were available in BCBaSe and used as a measure of SES. Expected mortality rates adjusted for education group were calculated using methods described in Study III so that the cumulative excess hazards could be estimated by education group.

6.4.2 Statistical methods

Multiple imputation

After creating a combined stage variable from data in BCBaSe, around 12% of this stage variable was missing. Additionally 3% of the education group information were missing. Falcaro *et al.* illustrated that results can be biased when calculating relative survival using a complete case analysis. Following their advice, we performed multinomial logistic regression using chained equations for the imputation of missing stage and education group [104]. This imputation model included age at diagnosis (RCS with three degrees of freedom), year of diagnosis (RCS with three degrees of freedom), region, educational group, the event indicator, and the Nelson-Aalen estimator of the cumulative hazard. We performed both a complete case analysis and an analysis where we had performed multiple imputation (MI) where 30 imputed values for each missing education group and stage value were created. Note that only the MI analysis is presented in Study IV.

Modelling approach and measures estimated

We performed a period analysis using the period window 2008-2012 [105]. Thus, instead of following all patients from their date of diagnosis until their date of death or censoring, we

left-truncated the follow-up times at 2008 and censored them at the end of 2012; person-time at risk only contributed to the analysis within this window. This method was chosen since it provides better future estimates than a standard cohort approach [106] since patients diagnosed recently contribute to the short term survival whereas long term survival is driven by patients diagnosed further into the past. We fitted flexible parametric excess hazard models to the breast cancer patients adjusting for age at diagnosis, stage at diagnosis and education group. From these models we estimated the LEL, postponable deaths and gain in life years that would have been expected under the following scenarios denoted methods A-C:

1. **Method A: What if we could remove differences in the stage distribution between education groups?** To answer this question we replaced the age group-specific stage distribution found in the low and the medium education groups with the age group-specific stage distribution seen in the high education group.

2. **Method B: What if we could remove differences in the breast cancer survival seen in different education groups?** We addressed this by replacing the age- and stage-specific relative survival found in the low and medium education groups with the age- and stage-specific relative survival seen in the high education group.

3. **Method C: What if we could remove differences between education groups both in terms of the stage distribution and the cancer-specific survival?** This was approached by combining methods A and B, and applying both the age-specific stage-distribution and the age- and stage-specific relative survival of the high education group to the low and medium education groups.

These three methods were used to estimate the effect of differences in educational group on survival differences in breast cancer patients; see also Figure 1 in Study IV for clarification using relative survival as an example. We applied these three different methods when estimating the LEL, gain in life years, total life years lost and postponable deaths.

Standardised estimates

We stage-standardised estimates of relative survival (used for calculation of postponable deaths) and estimates of the LEL. This was done in the following way for relative survival, and similarly for LEL:

$$\mathrm{R}_{e,a}^{\text{stage-stand}}(t) = \sum_{s=1}^{4} w_{e,a,s} \cdot \mathrm{R}_{e,a,s}(t) \tag{6.7}$$

where $w_{e,a,s}$ represented weights corresponding to the stage distribution in each age and education group. This was calculated as $w_{e,a,s} = n_{e,a,s}/n_{e,a}$ where n was the number of observations in each $e = 1,2,3$ education groups, $a = 1,...,6$ age groups, and $s = 1,2,3,4$ stage groups. This internal standardisation enabled us to provide one summary estimate of relative survival for each education group (<40 years, 40-49, 50-59, 60-69, 70-79, and 80+ years) and age whilst accounting for stage. Although age was categorised for the calculation of weights, age

was otherwise used as a continuous variable, i.e., predictions of LEL were calculated for each age separately. The total life years lost, total gain in life years and postponable deaths were calculated by summing over (continuous) age and by using the number of individuals at risk at each age. See the following section for further information on how these measures were calculated.

Postponable deaths and total life years lost

Stage-standardised estimates of relative survival and LEL as calculated in equation (6.7) were obtained for ages 21 to 99 years. We used the stage-standardised relative survival estimates to calculate the postponable deaths (PD) for each education group, e, as follows:

$$\mathrm{PD}_e(t) = \sum_{a=21}^{99} N_{e,a} \cdot \left\{ [\overline{\mathrm{R}}_{e,a}(t) - \mathrm{R}_{e,a}(t)] \cdot S^*_{e,a}(t) \right\} \tag{6.8}$$

where a is age, the expected survival $S^*_{e,a}(t)$ was also matched for sex and calendar year, $\mathrm{RS}_{e,a}(t)$ is the stage-standardised relative survival, and $\overline{\mathrm{RS}}_{e,a}(t)$ is the stage-standardised relative survival using one of the alternative methods A, B, or C.

Stage-standardised LEL was used in the following way to estimate the total gain in life years:

$$\text{Total gain in life years} = \sum_{e=1}^{2} \sum_{a=21}^{99} N_{e,a} \cdot (\mathrm{LEL}_{a,e} - \overline{\mathrm{LEL}}_{a,e}) \tag{6.9}$$

where $\mathrm{LEL}_{a,e}$ is the stage standardised LEL for age, a, and education group, e, and $\overline{\mathrm{LEL}}_{a,e}$ is the alternative LEL using either method A, B, or C. Note that gain in life years, in comparison to the total gain in life years, were calculated as $\mathrm{LEL}_{a,e} - \overline{\mathrm{LEL}}_{a,e}$ and presented by age and education group.

Confidence intervals for estimates

It is common to combine uncertainty of estimates using Rubin's Rules when performing MI. Current literature advises to combine estimates using Rubin's Rules on the scale in which it was modelled. For example, if we were interested in estimating survival but did this by modelling the log hazard function, we would combine the imputation-specific estimates on the log hazard scale and then translate this back to the survival scale [107]. To our knowledge there is no advice on how to combine standardised estimates such as those used in this study. Consequently, we instead used a bootstrapping method in the aim of combining the uncertainty involved in the standardised estimates and the MI. The bootstrap was implemented in the following way:

1. Impute missing stage and education group 30 times using the MI methods previously described.

2. Create one data set containing 30 variables for both stage and education group which correspond to the 30 imputed values calculated from the MI.

3. Bootstrap this dataset by drawing samples with replacement.

4. For each bootstrapped sample, loop over the 30 imputed variables and estimate each standardised measure of interest using the imputed variables separately. Average the 30 estimates of interest.

5. Save each of these averaged estimates of interest from each of the bootstrapped samples.

6. Repeat the bootstrapping 100 times. Use the 100 averaged estimates and the normal approximation method to estimate the confidence intervals.

6.4.3 Main results

Results indicated that on the population level, a gain of approximately 573 life years could be obtained for a typical calendar year if the stage-distribution in all education groups was the same as the high education group. When equating the stage-specific relative survival between education groups, it was estimated that approximately 692 life years could be gained in a typical calendar year. We estimated that approximately 50 deaths could be postponed five-years after diagnosis during a typical calendar year if it was possible to remove both stage-distribution and stage-specific relative survival differences between education groups. Approximately 25 deaths could be postponed five years beyond diagnosis if it was possible to only remove the stage-distribution differences and approximately 27 deaths could be postponed five years beyond diagnosis if differences in stage-specific relative survival could be removed.

6.4.4 Additional results: proportion of expected life lost

In addition to the results presented in the paper, we also calculated the proportion of expected life lost (PELL); see Table 6.5. The estimates of PELL predicted from models using the imputed data set indicate that younger patients lose a larger proportion of their remaining life years than older patients. For example a woman diagnosed with breast cancer at age 45 in the low education group is predicted to lose 25% of their remaining life whereas a similar woman diagnosed at age 75 would lose approximately 15% of their remaining life years due to their diagnosis of breast cancer. The PELL results also indicate that patients in the low education groups generally lose more than those in the medium or high education groups. We see that larger benefits, i.e. smaller proportions, are seen when considering method C and removing all differences in stage-distribution and stage-specific breast cancer survival between education groups than just removing one of the two as in methods A and B.

		Proportion of life years lost			
Age	**Education**	**Original**	**Method A**	**Method B**	**Method C**
45	Low	0.252	0.222	0.231	0.202
	Medium	0.238	0.228	0.217	0.207
	High	0.213	-	-	-
55	Low	0.181	0.158	0.164	0.143
	Medium	0.168	0.154	0.160	0.147
	High	0.152	-	-	-
65	Low	0.132	0.125	0.119	0.113
	Medium	0.120	0.118	0.118	0.116
	High	0.120	-	-	-
75	Low	0.171	0.145	0.155	0.131
	Medium	0.134	0.137	0.132	0.135
	High	0.139	-	-	-

Table 6.5: The proportion of life years lost due to breast cancer for patients diagnosed in three regions of Sweden between 2008 and 2012. 'Original' refers to predictions from models using imputed datasets but without applying methods A-C. Method **A** refers to predictions where differences in stage distribution between education groups are removed, Method **B** refers to estimates when removing differences in stage-specific relative survival between education groups, and Method **C** estimates when removing both differences in stage-distribution and stage-specific relative survival.

6.4.5 Sensitivity analyses

MI of missing stage resulted in an average stage distribution of 45.3%, 37.9%, 13.6% and 3.3% of individuals being diagnosed with breast cancer at stage I, II, III and IV, respectively. We performed sensitivity analyses to determine the conclusions that could have been drawn had we chosen to perform a complete case analysis; results are shown when applying method A in Table 6.6.

Larger differences in the life years lost between the MI and complete case analysis are seen in the younger and older patients. The complete case analysis estimated a larger number life years lost in the younger patients, and a smaller number of life years lost in the older patients in comparison to the imputed analysis. The largest difference of approximately half a year was seen for those aged 45 years. At the same time the gain in life years were largely unchanged between the imputed and complete case analyses. This indicates that it makes little difference whether a complete case analysis or MI is undertaken when altering the stage-distribution using method A. The difference in the life years lost between the two analyses stem from the differences in the population. These differences are consistent between the LEL and LEL when applying method A, resulting in a similar gain in life years for the MI and complete case analyses.

Age	Education	Life years lost (95% CI) MI	Life years lost (95% CI) CC	Gain in life years (95% CI) MI	Gain in life years (95% CI) CC
45	Low	8.68 (7.27, 10.09)	9.13 (7.60, 10.66)	1.21 (1.13, 1.30)	1.22 (1.12, 1.32)
	Medium	9.27 (8.34, 10.21)	9.83 (8.87, 10.80)	0.42 (0.38, 0.46)	0.41 (0.35, 0.47)
55	Low	4.75 (4.18, 5.32)	4.84 (425, 5.43)	0.67 (0.64, 0.71)	0.64 (0.60, 0.68)
	Medium	4.83 (4.35, 5.31)	5.05 (4.56, 5.54)	0.44 (0.40, 0.47)	0.45 (0.41, 0.49)
65	Low	2.66 (2.38, 2.93)	2.53 (2.26, 2.80)	0.15 (0.14, 0.17)	0.11 (0.10, 0.12)
	Medium	2.64 (2.40, 2.89)	2.61 (2.33, 2.89)	0.04 (0.03, 0.05)	0.02 (0.01, 0.03)
75	Low	1.93 (1.77, 2.08)	1.74 (1.60, 1.88)	0.35 (0.34, 0.37)	0.31 (0.29, 0.33)
	Medium	1.94 (1.78, 2.10)	1.79 (1.62, 1.97)	-0.04 (-0.06, -0.03)	-0.06 (-0.07, -0.05)

Table 6.6: LEL and gain in life years for breast cancer patients in Sweden when removing differences in stage-distribution between education groups; MI versus complete case (CC)

6.4.6 Discussion

It is important to consider what effects we attempt to capture when quantifying differences between education groups in the described ways. Screening attendance may differ by educational achievement which could contribute to the differences between education groups in stage at diagnosis. Health awareness and subsequent healthcare-seeking behaviour, which differs over SES groups, could have a similar effect. Differences in stage-specific survival could be explained by lifestyle differences, access to healthcare and even treatment differences. We can really only speculate which factors contribute to these differences and the potential interventions that could be implemented to improve survival. It is practically impossible to remove all of these differences with one intervention, but these results illustrate potential gains. Further work and resources would be required to quantify why these differences occur, but understanding the magnitude of these differences as shown in this study provide information to help prioritise further research and potential interventions.

For this application we found that MI did not make a large difference in terms of conclusions drawn. Nevertheless, imputation of stage can be highly important since missing stage information is often considered not missing completely at random. Missing stage is more commonly seen in older patients and patients with more advanced stage tumours since investigation of such tumours is often deemed unnecessary since such patients are likely to receive palliative treatment [108]. Moreover, inclusion of all individuals is highly important when wishing to calculating values on a population level such as the total life years lost or the total gain in life years.

The combination of estimating standardised measures and MI was methodologically com-

plex; it was not clear how to undertake such a project. Regularly, Rubin's Rules could be used to estimate uncertainty in estimates calculated using MI, however, combining using Rubin's Rules was not an approach we considered here due to the standardisation. We instead used a bootstrap approach which was 1) specific to this application, and 2) computationally intensive. For these reasons, it would be useful to investigate generalised approaches for estimating uncertainty in standardised measures estimated using imputed data.

The results presented would have been more useful if we had access to data from all regions of Sweden, since we were truly interested in the effects for the whole country. We did present estimates of the total life years lost and postponable deaths for the whole of Sweden, but in order to do this we assumed that the study population represented Sweden in terms of expected survival, relative survival, covariate effects and the stage distribution over education groups. Under these assumptions we estimated that 1,340 life years could be saved for one calendar year if stage-distributions could be equated over education groups (method A), and 2,878 life years if stage-distributions and stage-specific relative survival could be equated across education groups (method C). Although these assumptions are quite strong, and we suggest that readers do not overinterpret these results, there are few reasons to believe that estimating over the whole of Sweden would lead to dramatic changes to these estimates. It would, however, be interesting to see how these assumptions affect the estimates if the data were available.

6.4.7 Conclusion

This study illustrates an alternative way to apply the LEL in quantifying survival of cancer patients. We use counter-factual situations to quantify the impact a characteristic has on the remaining life expectancy, the life years lost and the postponable deaths for a whole population. Such estimates quantify the potential gains interventions could bring if implemented. In addition to gaining a better understanding of the impact of a factor on prognosis, this sort of analysis could be used in a more practical sense to help determine what healthcare interventions are worth investing in. We conclude that removing differences in both the stage distribution at diagnosis of breast cancer and the stage-specific relative survival between education groups could postpone approximately 50 deaths five years beyond diagnosis per year in three regions of Sweden. On a national level, an estimated 117 deaths could be postponed five years beyond diagnosis per year.

7. Overall conclusions and future perspectives

Study I, whose results apply even outside of the population-based cancer studies field, showed that non-proportional FPMs can be used to estimate survival, hazard rates and HRs with low levels of bias when the selection of the degrees of freedom are guided by the AIC and BIC. Users should perform relevant sensitivity analyses and consider that model mis-specification can affect predicted hazard rates and ratios more than survival proportions.

Studies II and IV illustrated how the LEL can provide a further understanding of cancer patient survival and works as a complement to the relative survival measure. Our application to CML in Study II showed how the LEL can be used to assess the impact treatment improvements can have on the life expectancy over time. We concluded that CML patients can now expect to live almost as long as CML-free individuals.

The LEL can also be used to address questions about the potential population impact of risk factors such as education. In Study IV we estimated LEL and postponable deaths in breast cancer patients and determined the effect of differences between education groups. Such results can help guide decisions on interventions aimed to remove survival differences between patient groups.

We concluded from Study III that information from a random sample of the general population can be utilised to estimate adjusted expected mortality rates using both Poisson models and FPMs. This work offers the opportunity for quantification of survival measures by non-standard factors which is an essential next step to understand remaining differences in survival between patients.

Overall, each of the studies address and add to slightly different aspects of methodology and understanding of survival of cancer patients in the population-based cancer studies field. The work in these areas can always be developed further; here are some thoughts on extensions and applications of the methods and results shown in this thesis.

Previous work has shown that extrapolation of relative survival is accurate in patients over 50 years of age, but there has been no assessment of the extrapolation for patients younger than this age. The LEL is a highly useful measure due to its interpretation in terms of life expectancy, and although life expectancy is of interest to everyone regardless of age, the impact of a cancer diagnosis for younger patients is much larger since they have a much longer remaining life expectancy and thus much more potential life to lose. This makes this measure

particularly relevant for younger patients. The problems with extrapolation in these younger patients include 1) extrapolation of relative survival in order to estimate LEL will be longer, and 2) for many cancers, diagnoses in younger patients is less common, thus there are fewer cases and increased uncertainty in any extrapolation. Work in this area would be of interest to ensure that LEL can be accurately estimated for younger patients, or to determine when LEL should not be estimated due to problems with extrapolation.

Similarly, applications of the LEL to other cancers is highly useful for the several reasons. The LEL provides an alternative aspect of cancer survival in comparison to the common practice of estimating and presenting relative survival. LEL can also be a helpful tool in understanderstanding what a diagnosis of cancer means for patients' future life. Cancer registries commonly report relative survival estimates; presenting such estimates is essential when comparing different subgroups and populations. Nevertheless, the inclusion of 'real-world' estimates, such as total life years lost, in reports from cancer registries may be advantageous to understand the burden cancer has on a population. Inclusion of such measures in reports from cancer registries or other public reports or databases such as the NORDCAN project [19] is extremely important in an age where the internet offers everyone access to information on cancer statistics. Patients are using information published online to try to understand their diagnosis of cancer and the likely impact it will have [109]. Presenting more easily understood measures such as the LEL means they are more likely to intepret survival accurately.

Online tools are becoming more popular and have been implemented to illustrate particular methodology via interactive graphs [110], and to help aid the understanding of cancer survival statistics [111]. Such online tools could be developed to enhance understanding of both FPMs and LEL estimates. Illustrating how changes in the degrees of freedom in FPMs alter commonly used predictions could be useful for users to understand what model specifications mean and where biases may occur. Online tools could also be useful to help understand the impact of certain modelling assumptions when estimating LEL.

Study III implemented methods to smooth expected mortality rates. It is interesting to consider the effects of smoothing expected mortality rates on measures such as relative survival and whether such smoothing is actually needed. Approaches for smoothing rates in abridged life tables seem sensible [65], but it is not clear how small the time intervals need to be before smoothing is unnecessary. It may be interesting to investigate the implications of either assuming expected rates are constant within specific time intervals versus the implications of making modelling assumptions.

Related to this, the expected mortality rates are commonly assumed to carry no uncertainty since they are obtained from a whole nation. One could alternatively follow the theory that observed population mortality rates are taken from some underlying random process. Previous research showed that accounting for this uncertainty does not significantly affect model-based relative survival estimates [102]. However, when estimating LEL expected mortality rates are used 1) to estimate relative survival, 2) to transform the relative survival back to the observed survival, and 3) in estimating the life expectancy that patients would be expected to experience had they not been diseased. Ignoring uncertainty in all three parts of estimating the LEL may

have a larger impact and it would be interesting to assess whether this is true.

In the same manner, when modelling excess mortality using FPMs and estimating LEL, we make one of two assumptions about the future expected mortality. We either 1) assume that future expected rates are the same as the most recently recorded year, or 2) use predictions estimated by demographers; in our applications predictions were obtained from Statistics Sweden. It is always problematic when trying to predict the future, and although we have performed some sensitivity analyses regarding these two assumptions in our specific applied studies, it may be useful to do more extensive work into the differences that these two assumptions can make and if one is preferable.

The MI in Study IV was undertaken using advice from Falcaro *et al.* on imputation using relative survival. However, other methods for imputation in relative survival seem to be rather limited. For example, if cause of death information is available, it could be useful to utilise this information. If we knew that an individual died of cancer a short time after diagnosis of that cancer, it seems more likely that they had a more advanced stage cancer at diagnosis. It would be interesting to perform a simulation study using such information to assess whether cause of death improves imputated data. It would also be useful to extend and generalise the methodology used to calculate uncertainty when combining MI with standardised-measures as performed in Study IV.

Further work related to FPMs in general could be to extend the use of FPMs on the log hazard scale to modelling multiple timescales. It may be desirable to capture multiple timescales in certain analyses, for example if we were interested in both time since diagnosis and attained age, i.e., aging. Whilst working on Study I and Study III, and the `strcs` command in Stata, it became clear that it was possible to model multiple timescales using FPMs on the log hazard scale. The use of the numerical integration to calculate the log likelihood allows us to model one timescale as a function of the other, since time progresses in the same way. This could be interesting work since it removes the computationally intensive problem of time-splitting that usually takes place when wishing to perform an analysis with multiple timescales.

A. Appendix

A.1 Hazard ratios estimated from flexible parametric models on the log cumulative hazard scale

As described in section 5.1, coefficients estimated in FPMs can be interpreted as log HRs provided there is at most one time-dependent effect. Consider the following FPM on the log cumulative hazard scale with one time-dependent effect:

$$\begin{aligned}\log[H(t)] &= s_0(\alpha|\gamma) + \beta_1 x + s_1(\alpha|\delta_1)x + \beta_2 z \\ \Rightarrow H(t) &= \exp[s_0(\alpha|\gamma) + \beta_1 x + s_1(\alpha|\delta_1)x + \beta_2 z],\end{aligned}$$

where $\alpha = \ln(t)$, x and z are indicator variables, and x has a time-dependent effect. If we set x to be any value, i.e., $x = x_i$, we can write the HR for $z = 1$ versus $z = 0$ as follows:

$$\begin{aligned}\frac{h(t|z=1, x=x_i)}{h(t|z=0, x=x_i)} &= \frac{d}{dt}\frac{H(t|z=1)}{H(t|z=0)} \\ &= \frac{d}{dt}\frac{\exp[s_0(\alpha|\gamma) + \beta_1 x_i + s_1(\alpha|\delta_1)x_i + \beta_2]}{\exp[s_0(\alpha|\gamma) + \beta_1 x_i + s_1(\alpha|\delta_1)x_i]} \\ &= \frac{d}{dt}\frac{\exp[s_0(\alpha|\gamma)]\exp(\beta_1 x_i)\exp[s_1(\alpha|\delta_1)x_i]\exp(\beta_2)}{\exp[s_0(\alpha|\gamma)]\exp(\beta_1 x_i)\exp[s_1(\alpha|\delta_1)x_i]} \\ &= \frac{d}{dt}\exp(\beta_2) \\ &= \exp(\beta_2)\end{aligned}$$

Thus β_2 can be interpreted as the log HR for z. When modelling on the log cumulative hazard scale and including more than one time-dependent effect, problems arise. These are described in the following section.

A.2 Modelling multiple time-dependent effects using flexible parametric models on the log cumulative hazard scale

Consider the following FPM with time-dependent effects of both x and z:

$$\begin{aligned}\log[H(t)] &= s_0(\alpha|\gamma) + \beta_1 x + s_1(\alpha|\delta_1)x + \beta_2 z + s_2(\alpha|\delta_2)z \\ \Rightarrow H(t) &= \exp[s_0(\alpha|\gamma) + \beta_1 x + s_1(\alpha|\delta_1)x + \beta_2 z + s_2(\alpha|\delta_2)z].\end{aligned}$$

Differentiating this in order to get the hazard function (note that for ease of reading RCS functions are now denoted s_k):

$$h(t) = (s_0' + s_1' x + s_2' z)\exp(s_0 + \beta_1 x + s_1 x + \beta_2 z + s_2 z).$$

Now, if we want the HR for $z = 1$ versus $z = 0$, where $x = x_i$, we get the following:

$$\begin{aligned}\frac{h(t|z=1, x=x_i)}{h(t|z=0, x=x_i)} &= \frac{(s_0' + s_1' x_i + s_2')\exp(s_0 + \beta_1 x_i + s_1 x_i + \beta_2 + s_2)}{(s_0' + s_1' x_i)\exp(s_0 + \beta_1 x_i + s_1 x_i)} \\ &= \frac{(s_0' + s_1' x_i + s_2')\exp(s_0)\exp(\beta_1 x_i)\exp(s_1 x_i)\exp(\beta_2)\exp(s_2)}{(s_0' + s_1' x_i)\exp(s_0)\exp(\beta_1 x_i)\exp(s_1 x_i)} \\ &= \frac{(s_0' + s_1' x_i + s_2')\exp(\beta_2)\exp(s_2)}{s_0' + s_1' x_i}\end{aligned}$$

The HR for z is still dependent on the differentiated spline term for x, thus illustrating the problem when modelling multiple time-dependent effects using FPMs on the log cumulative scale. It should be noted that modelling with multiple time-dependent effects on the log cumulative hazard scale is only a large problem when one is interested in reporting HRs. This was not a problem for the studies presented in this thesis as focus was on individual level predictions, not on HRs.

Acknowledgements

It feels like there are an endless number of people to acknowledge for their support, inspiration and friendship throughout this journey. Here's my short attempt at thanking you all.

Thank you to my main supervisor **Paul Lambert** for your excellent supervision and guidance over the past four years. You have given me the opportunity to make my own mistakes and create my own successes, and have always been there to solve problems before everything fell apart. Thank you for your patience, positive attitude to research, and for believing in me!

To my co-supervisors **Paul Dickman**, **Mats Lambe** and **Therese Andersson**. Thank you **Paul** for always being open to discussions, for asking difficult questions, and for teaching me that the most important thing in research is to have fun! Thanks to **Mats** for taking the time to discuss our research and for sharing your knowledge. Thank you **Therese** for being a friend as well as a supervisor; you have given me a huge amount of support and been there for me in every aspect of my PhD for which I am grateful.

Anna Johansson and **Caroline Weibull**: thank you for all your help over the years and for inspiring me through your honesty, enthusiasm and hardworking attitude.

To the other members of the group, past and present: **Michael Crowther**, **Sandra Eloranta**, **Sarwar Islam**, **Mark Rutherford**, and **Betty Syriopoulou**. Thanks for all your help, our discussions, and especially the laughs!

To my other collaborators **Magnus Björkholm**, **Mark Clements**, **Martin Höglund**, and **Xingrong Liu**. Thank you for your contribution to this work, I am grateful to you for sharing your insight and knowledge.

Thank you to my mentor **Thomas Frisell** for imparting some of your research wisdom onto me and for complementing my PhD studies with an education in fine dining, wine and class.

To the **MEB biostatisticians**: thank you for teaching me so much, and making me realise there is still a great deal left to learn.

Marie Jansson, **Camilla Ahlqvist**, **Gunilla Sonnebring** and other friendly faces at MEB: thank you for being incredibly helpful and making every MEBbers life a little bit brighter.

To the PhD students and other researchers at MEB including, but not limited to, **Isabell Brikell**, **Bénédicte Delcoigne**, **Malin Ericsson**, **Laura Ghirardi**, **Nelson Ndegwa Gichora**, **Johanna Holm**, **Andreas Jangmo**, **Andreas Karlsson**, **Ida Karlsson**, **Favelle Lamb**, **Elisa Longinetti**, **Frida Lundberg**, **Daniela Mariosa**, **Emilio Ugalde Morales**, **Zheng Ning**, **Tobba Palsdottir**, **Anna Plym**, **Cecilia Radkiewicz**, **Vilhelmina Ullemar**, and **Shuyang Yao**. So many of you have inspired me throughout this journey with your dedicated, hardworking spirit. Thank you!

A special shout out to **Kathleen Bokenberger**, **Elisabeth Dahlqwist**, **Gabriel Isheden**, and **Johan Zetterqvist**. You guys have really made my experience at MEB a great one. Major thanks to **Kat** for all the honest discussions, for your sense of humour, and for keeping me sane.

To all my amazing friends, wherever you may be: thank you for your support, friendship and the confidence boosts. Thanks to **Beck**, **Sarah**, and **Johno**, because no one has ever made me laugh as hard as you three fools over all these years. To the **Leicester ladies**: I wouldn't have got this far without you! Thanks to my ever changing and growing friendship group in Stockholm, especially **Pilar**, **Fernando**, and **Jack** who've been there with me from the beginning of my PhD adventure. A PhD student needs to let off some steam sometimes and you guys have been the perfect team for all such things.

Thank you **Andy**, **James**, **Tom** and the rest of my family for your love and constant encouragement. Thank you to my brother **Luke** for keeping me grounded and always looking out for me, and to my niece **Mia** for always making me smile. A special thanks to my **Mum** who has always helped me get through stressful times; thank you for all you have done for me, especially for forcing me to get on that plane to Sweden.

A massive thanks to **Ross** for continually telling me I'm good enough to do this. For encouraging me to take risks and comforting me in my failures. If just for your sake, I'll try to laugh more and moan less in my future endeavours.

And finally, to my **Dad**, my number one inspiration: this one is for you.

References

[1] The National Board of Health and Welfare (Socialstyrelsen). Statistics on Cancer Incidence 2015 [Art. no. 2017-1-20]; 2017.

[2] The National Board of Health and Welfare (Socialstyrelsen). Statistics on Causes of Death 2015 [Art. no. 2016-8-4]; 2015.

[3] Armstrong BK. The Role of the Cancer Registry in Cancer Control. Cancer Cause Control. 1992;3(6):569–579.

[4] Maxwell DM. The role of cancer registries in cancer control. Int J Clin Oncol. 2008;13(2):102–111.

[5] Dickman PW, Adami HO. Interpreting trends in cancer patient survival. J Intern Med. 2006;260(2):103–117.

[6] Ellis L, Woods LM, Estève J, Eloranta S, Coleman MP, Rachet B. Cancer incidence, survival and mortality: explaining the concepts. Int J Cancer. 2014;135(8):1774–1782.

[7] O'Connor CM, Adams JU. Essentials of Cell Biology. Cambridge, MA: NPG Education; 2010.

[8] Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57–70.

[9] Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Molecular Cell Biology, 4th edition. New York: W. H. Freeman; 2000.

[10] National Cancer Institute. Risk Factors for Cancer [Internet]; 2015 [Accessed Nov 21, 2017]. Available from: `https://www.cancer.gov/about-cancer/causes-prevention/risk`.

[11] Björkholm M, Ohm L, Eloranta S, Derolf A, Hultcrantz M, Sjöberg J, et al. Success story of targeted therapy in chronic myeloid leukemia: a population-based study of patients diagnosed in Sweden from 1973 to 2008. J Clin Oncol. 2011;29(18):2514–2520.

[12] Olsson-Strömberg U, Simonsson B, Ahlgren T, Björkholm M, Carlsson K, Gahrton G, et al. Comparison of busulphan, hydroxyurea and allogeneic bone marrow transplantation (BMT) in chronic myeloid leukaemia: BMT prolongs survival. The Hematology Journal. 2004;5(6):462–6.

[13] O'Brien SG, Guilhot F, Larson RA, Gathmann I, Baccarani M, Cervantes F, et al. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. New England Journal of Medicine. 2003;348(11):994–1004.

[14] Mahon FX. Treatment-free remission in CML: who, how, and why? Hematology. 2017;2017(1):102–109.

[15] Höglund M, Sandin F, Simonsson B. Epidemiology of chronic myeloid leukaemia: an update. Annals of Hematology. 2015;94(Suppl 2):S241–S247.

[16] Gunnarsson N, Sandin F, Höglund M, Stenke L, Björkholm M, Lambe M, et al. Population-based assessment of chronic myeloid leukemia in Sweden: striking increase in survival and prevalence. European Journal of Haematology. 2016;97(4):387–92.

[17] The National Board of Health and Welfare (Socialstyrelsen). Cancer Incidence in Sweden 2014 [Art. no. 2014-12-26]; 2015.

[18] The National Board of Health and Welfare (Socialstyrelsen). Cancer Incidence in Sweden 2011 [Art. no. 2012-12-19]; 2012.

[19] Engholm G, Ferlay J, Christensen N, Kejs AMT, Hertzum-Larsen R, Johannesen TB, et al.. NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries. Version 7.3 [Internet]; 2016 [Accessed Nov 24, 2017]. Available from: `http://www.ancr.nu`.

[20] Colzani E. Health Outcomes of Women with Breast Cancer [PhD Thesis]. Stockholm: Karolinska Institutet; 2014 [Accessed Jan 8, 2018]. Available from: `https://openarchive.ki.se/xmlui/handle/10616/42321`.

[21] Lagerlund M, Bellocco R, Karlsson P, Tejler G, Lambe M. Socio-economic factors and breast cancer survival - a population-based cohort study (Sweden). Cancer Causes and Control. 2005;16(4):419–430.

[22] Colzani E, Liljegren A, Johansson ALV, Adolfsson J, Hellborg H, Hall PFL, et al. Prognosis of patients with breast cancer: causes of death and effects of time since diagnosis, age, and tumor characteristics. J Clin Oncol. 2011;29(30):4014–4021.

[23] Hussain SK, Altieri A, Sundquist J, Hemminki K. Influence of education level on breast cancer risk and survival in Sweden between 1990 and 2004. International Journal of Cancer. 2008;122(1):165–169.

[24] Downing A, Prakash K, Gilthorpe MS, Mikeljevic JS, Forman D. Socioeconomic background in relation to stage at diagnosis, treatment and survival in women with breast cancer. British journal of cancer. 2007;96(5):836–840.

[25] Lundqvist A, Andersson E, Ahlberg I, Nilbert M, Gerdtham U. Socioeconomic inequalities in breast cancer incidence and mortality in Europe: a systematic review and meta-analysis. The European Journal of Public Health. 2016;26(5):804–813.

[26] Leung KM, Elashoff RM, Afifi AA. Censoring Issues in Survival Analysis. Annual Reviews in Public Health. 1997;18:83–104.

[27] Austin P. A Tutorial on Multilevel Survival Analysis:Methods, Models and Applications. International Statistical Review. 2017;85(2):185–203.

[28] Cox DR. Regression models and life-tables. JRSSB. 1972;34(2):187–220.

[29] Ryan TP, Woodall WH. The most-cited statistical papers. Journal of Applied Statistics. 2005;32(5):461–474.

[30] Carstensen B. Age-period-cohort models for the Lexis diagram. Statistics in Medicine. 2007 Jul;26(15):3018–3045.

[31] Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. Annu Rev Public Health. 1999;20:145–57.

[32] Ludvigsson JF, Almqvist C, Bonamy AK, Ljung R, Michaëlsson K, Neovius M, et al. Registers of the Swedish total population and their use in medical research. European Journal of Epidemiology. 2016;31(2):125–36.

[33] The National Board of Health and Welfare (Socialstyrelsen). Swedish Cancer Registry [Internet]; [Accessed December 11, 2017]. Available from: `http://www.socialstyrelsen.se/register/halsodataregister/cancerregistret/inenglish`.

[34] Barlow L, Westergren K, Holmberg L, Talbäck M. The completeness of the Swedish Cancer Register: a sample survey for year 1998. Acta Oncol. 2009;48(1):27–33.

[35] The Human Mortality Database [Internet]. University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany); [Accessed Dec 11, 2017]. Available from: `http://www.mortality.org/` or `http://www.humanmortality.de/`.

[36] Royston P, Lambert PC. Flexible parametric survival analysis in Stata: Beyond the Cox model. First edition ed. Texas: Stata Press; 2011.

[37] Royston P. Flexible parametric alternatives to the Cox model, and more. The Stata Journal. 2001;1(1):1–28.

[38] Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. Stat Med. 2007;26(30):5486–5498.

[39] Andersson TML, Dickman PW, Eloranta S, Lambert PC. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. BMC Med Res Methodol. 2011;11:96.

[40] Crowther MJ, Look MP, Riley RD. Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. Stat Med. 2014;33(22):3844–3858.

[41] Durrleman S, Simon R. Flexible regression models with cubic splines. Statistics in medicine. 1989;8(5):551–561.

[42] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. The Stata Journal. 2009;9(2):265–290.

[43] Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. Journal of Statistical Computation and Simulation. 2015;85(4):777–793.

[44] Stata Corporation. Stata Statistical Software, Release 15 [computer software]. College Station, TX; 2017. Available from: `https://www.stata.com/`.

[45] Bower H, Crowther MJ, Lambert PC. strcs: A command for fitting flexible parametric survival models on the log-hazard scale. The Stata Journal. 2016;16(4):989–1012.

[46] Crowther MJ, Lambert PC. stgenreg: A Stata Package for General Parametric Survival Analysis. Journal of Statistical Software. 2013;53:1–17.

[47] Akaike H. Information theory as an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second International Symposium on Information Theory. Budapest: Akademiai Kiado; 1973. p. 267–281.

[48] Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978;6(2):461–464.

[49] Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. Biometrics. 2000;56(1):256–262.

[50] Dickman PW, Coviello E. Estimating and modelling relative survival. The Stata Journal. 2015;15(1):186–215.

[51] Ederer F, Axtell LM, Cutler SJ. The relative survival rate: A statistical methodology. National Cancer Institute Monograph. 1961;6:101–121.

[52] Ji J, Sundquist K, Sundquist J, Hemminki K. Comparability of cancer identification among Death Registry, Cancer Registry and Hospital Discharge Registry. Int J Cancer. 2012;131(9):2085–2093.

[53] Corazziari I, Quinn M, Capocaccia R. Standard cancer patient population for age standardising survival ratios. Eur J Cancer. 2004 Oct;40(15):2307–2316.

[54] Ederer F, Heise H. Instructions to IBM 650 programmers in processing survival computations. Methodological note No. 10, End Results Evaluation Section. National Cancer Institute, Bethesda, MD; 1959.

[55] Pohar Perme M, Stare J, Estève J. On Estimation in Relative Survival. Biometrics. 2012;68(1):113–120.

[56] Lambert PC, Dickman PW, Rutherford MJ. Comparison of approaches to estimating age-standardized net survival. BMC Med Res Methodol. 2015;15:64.

[57] Seppä K, Hakulinen T, Läärä E, Pitkäniemi J. Comparing net survival estimators of cancer patients. Stat Med. 2016;35(11):1866–1879.

[58] Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. Statistics in Medicine. 1990;9(5):529–538.

[59] Berkson J. The calculation of survival rates. In: Walters W, Gray HK, Priestly JT, editors. Carcinoma and other malignant lesions of the stomach. Philadelphia: Sanders; 1942. p. 467–484.

[60] Grafféo N, Jooste V, Giorgi R. The impact of additional life-table variables on excess mortality. Statistics in Medicine. 2012;31:4219–30.

[61] Blakely T, Soeberg M, Carter K, Costilla R, Atkinson J, Sarfati D. Bias in relative survival methods when using incorrect life-tables: lung and bladder cancer by smoking status and ethnicity in New Zealand. Int J Cancer. 2012;131(6):E974–E982.

[62] Dickman PW, Auvinen A, Voutilainen ET, Hakulinen T. Measuring social class differences in cancer patient survival: is it necessary to control for social class differences in general population mortality? A Finnish population-based study. J Epidemiol Community Health. 1998;52(11):727–734.

[63] Jeffreys M, Stevanovic V, Tobias M, Lewis C, Ellison-Loschmann L, Pearce N, et al. Ethnic inequalities in cancer survival in New Zealand: linkage study. American Journal of Public Health. 2005;95:834–7.

[64] Eloranta S, Lambert PC, Cavalli-Björkman N, Andersson TML, Glimelius B, Dickman PW. Does socioeconomic status influence the prospect of cure from colon cancer–a population-based study in Sweden 1965-2000. Eur J Cancer. 2010 Nov;46(16):2965–2972.

[65] Rachet B, Maringe C, Woods LM, Ellis L, Spika D, Allemani C. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. BMC Public Health. 2015;15:1240.

[66] Ellis L, Coleman MP, Rachet B. The impact of life tables adjusted for smoking on the socio-economic difference in net survival for laryngeal and lung cancer. Br J Cancer. 2014 Jul;111(1):195–202.

[67] Morris M, Woods LM, Rachet B. A novel ecological methodology for constructing ethnic-majority life tables in the absence of individual ethnicity information. Journal of epidemiology and community health. 2015;69:361–7.

[68] Howlader N, Ries LAG, Mariotto AB, Reichman ME, Ruhl J, Cronin KA. Improved estimates of cancer-specific survival rates from population-based data. J Natl Cancer Inst. 2010 Oct;102(20):1584–1598.

[69] Talbäck M, Dickman PW. Estimating expected survival probabilities for relative survival analysis - Exploring the impact of including cancer patient mortality in the calculations. Eur J Cancer. 2011;47(17):2626–32.

[70] Andersson TML, Dickman PW, Eloranta S, Lambe M, Lambert PC. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. Statistics in Medicine. 2013;32:5286–5300.

[71] Hakama M, Hakulinen T. Estimating the expectation of life in cancer survival studies with incomplete follow-up information. Journal of Chronic Diseases. 1977;30(9):585–597.

[72] Andersson TML, Lambert PC. Fitting and modeling cure in population-based cancer studies within the framework of flexible parametric survival models. The Stata Journal. 2012;12(4):623–628.

[73] Rutherford MJ, Andersson TML, Møller H, Lambert PC. Understanding the impact of socioeconomic differences in breast cancer survival in England and Wales: avoidable deaths and potential gain in expectation of life. Cancer Epidemiology. 2015;39(1):118–125.

[74] Bower H, Andersson TML, Bjorkholm M, Dickman PW, Lambert PC, Derolf AR. Continued improvement in survival of acute myeloid leukemia patients: an application of the loss in expectation of life. Blood Cancer Journal. 2016;6:e390.

[75] Baade PD, Youlden DR, Andersson TML, Youl PH, Kimlin MG, Aitken JF, et al. Estimating the change in life expectancy after a diagnosis of cancer among the Australian population. BMJ Open. 2015;5:e006740.

[76] Andersson TML, Dickman PW, Eloranta S, Sjövall A, Lambe M, Lambert PC. The loss in expectation of life after colon cancer: a population-based study. BMC Cancer. 2015;15(1):412.

[77] Nelson CP, Lambert PC, Squire IB, Jones DR. Relative survival: what can cardiovascular disease learn from cancer? European Heart Journal. 2008;29(7):941–947.

[78] Andersen PK. Life years lost among patients with a given disease. Statistics in Medicine. 2016;36(22):3573–3582.

[79] Burnet NG, Jefferies SJ, Benson RJ, Hunt DP, Treasure FP. Years of life lost (YLL) from cancer is an important measure of population burden – and should be considered when allocating research funds. Br J Cancer. 2005;92(2):241–245.

[80] Brustugun OT, Moller B, Helland A. Years of life lost as a measure of cancer burden on a national level. Br J Cancer. 2014;111:1014–20.

[81] The World Health Organisation. Years of life lost (percentage of total) [Internet]; 2006 [Accessed Dec 21, 2017]. Available from: `http://www.who.int/whosis/whostat2006YearsOfLifeLost.pdf`.

[82] Mettlin C. Trends in years of life lost to cancer: 1970-1985. CA Cancer J Clin. 1989;39(1):33–39.

[83] Horm JW, Sondik EJ. Person-years of life lost due to cancer in the United States, 1970 and 1984. Am J Public Health. 1989;79(11):1490–1493.

[84] Andersson T. Quantifying cancer patient survival: extensions and applications of cure models and life expectancy estimation [PhD Thesis]. Stockholm: Karolinska Institutet; 2013. Available from: `https://openarchive.ki.se/xmlui/handle/10616/41668`.

[85] Syriopoulou E, Bower H, Andersson TML, Lambert PC, Rutherford MJ. Estimating the impact of a cancer diagnosis on life expectancy by socio-economic group for a range of cancer types in England. Epidemiology. 2017;117(9):1419–1426.

[86] Rutherford MJ, Ironmonger L, Ormiston-Smith N, Abel GA, Greenberg DC, Lyratzopoulos G, et al. Estimating the potential survival gains by eliminating socioeconomic and sex inequalities in stage at diagnosis of melanoma. British Journal of Cancer. 2015;112 Suppl:S116–S123.

[87] R B, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine. 2005;24(11):1713–1723.

[88] Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Statistics in Medicine. 2013;32(23):4118–4134.

[89] Coleman MP, Babb P, Damiecki P, Grosclaude P, Honjo S, Jones J, et al. Cancer Survival Trends in England and Wales, 1971–1995: Deprivation and NHS Region. No. 61 in Studies in Medical and Population Subjects. London: The Stationery Office; 1999.

[90] Lambert PC, Dickman PW, Österlund P, Andersson TML, Sankila R, Glimelius B. Temporal trends in the proportion cured for cancer of the colon and rectum: a population-based study using data from the Finnish Cancer Registry. International Journal of Cancer. 2007;121(9):2052–2059.

[91] Abrahamowicz M, MacKenzie T, M EJ. Time-dependent hazard ratio: modeling and hypothesis testing wit5h application in lupus nephritis. Journal of the American Statistical Association. 1996;91(436):1432–1439.

[92] Gu C. Model diagnostics for smoothing spline ANOVA models. Canadian Journal of Statistics- Revue Canadienne de Statistique. 2004;32(4):347–358.

[93] Pawitan Y. In All Likelihood: Statistical Modelling and Inference Using Likelihood. First edition ed. Oxford: Oxford University Press; 2001.

[94] Kullback S, Leibler RA. On Information and Sufficiency. Annals of Mathematical Statistics. 1951;22(1):79–86.

[95] Emiliano PC, Vivanco MJF, de Menezes FS. Information criteria: How do they behave in different models? Computational Statistics & Data Analysis. 2014;69(0):141 – 153.

[96] Liu XR, Pawitan Y, Clements M. Parametric and penalized generalized survival models. Statistical Methods in Medical Research. 2016;0(0):1–16. DOI:10.1177/0962280216664760.

[97] Gunnarsson N, Stenke L, Höglund M, Sandin F, Björkholm M, Dreimane A, et al. Second malignancies following treatment of chronic myeloid leukaemia in the tyrosine kinase inhibitor era. British Journal of Haematology. 2015;169(5):683–688.

[98] Dahlén T, Edgren G, Lambe M, , Höglund M, Björkholm M, et al. Cardiovascular Events Associated With Use of Tyrosine Kinase Inhibitors in Chronic Myeloid Leukemia: A Population-Based Cohort Study. Annals of Internal Medicine. 2016;165(3):161–6.

[99] Baccarani M, Pane F, Rosti G, Russo D, Saglio G. Chronic myeloid leukemia: room for improvement? Haematologica. 2017;102(7):1131–1133.

[100] Pulte D, Jansen L. Population-Level Survival for Patients With Chronic Myeloid Leukemia: Higher Survival in Sweden Than Internationally. Journal of Clinical Oncology. 2016;35(6):695–6.

[101] Etienne G, Guilhot J, Rea D, Rigal-Huguet F, Nicolini F, Charbonnier A, et al. Long-Term Follow-Up of the French Stop Imatinib (STIM1) Study in Patients With Chronic Myeloid Leukemia. Journal of Clinical Oncology. 2017;35(3):298–305. PMID: 28095277.

[102] Gauffin O. Confidence Intervals in Relative Survival Analysis [Masters degree project]. Stockholm: Stockholm University; 2017.

[103] Rubin DB. Multiple Imputation for Nonresponse in Surveys. New Jersey: John Wiley and Sons; 2004.

[104] Falcaro M, Nur U, Rachet B, Carpenter JR. Estimating excess hazard ratios and net survival when covariate data are missing: strategies for multiple imputation. Epidemiology. 2015;26(3):421–428.

[105] Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. Cancer. 1996;78(9):2004–2010. DOI:10.1186/1471-2288-9-57.

[106] Brenner H, Hakulinen T. Up-to-date cancer survival: period analysis and beyond. Int J Cancer. 2009;124(6):1384–1390.

[107] Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Medical Research Methodology. 2009;9:1–8.

[108] Nur U, Shack LG, Rachet B, Carpenter JR, Coleman MP. Modelling relative survival in the presence of incomplete data: a tutorial. International Journal of Epidemiology. 2009;39(1):118 –128.

[109] Carlsson ME. Cancer patients seeking information from sources outside the health care system: change over a decade. Eur J Oncol Nurs. 2009;13(4):304–305.

[110] Lambert PC. Interactive Graphs [Internet] University of Leicester; [Accessed December 21, 2017]. Available from: `https://www2.le.ac.uk/Members/pl4/interactive-graphs`.

[111] Islam Mozumder S, Rutherford MJ, Dickman PW, Lambert PC. InterPreT: Interactive cancer survival prediction tool; 2016 [Accessed Nov 27, 2017]. Available from `https://interpret.le.ac.uk/index.php`.